

УДК 811.161

О.М.Демська-Кульчицька

ПАРАДИГМА МЕТАДАНИХ
У ЗАГАЛЬНОМОВНОМУ УКРАЇНСЬКОМУ
КОРПУСІ ТЕКСТІВ

Сучасна лінгвістична наука є внутрішньо здиференційованою системою окремих, пов'язаних як між собою, так і з іншими науками, галузей і напрямів, серед яких важливе місце посідає напрям **корпусного мовознавства**, що виник у 60-х роках минулого століття та інтенсивно розвивається впродовж останніх років минулого і початку нашого століття.

Об'єктом дослідження корпусної лінгвістики є тексти природної мови, організовані як **корпус текстів** – машиночитане, стандартно подане зібрання репрезентативних для певної мови, діалекту або іншої підмножини мов писемних або усних текстів, призначене для лінгвального аналізу та лінгвістичного опису, відібраних і впорядкованих згідно з експліцитними лінгвістичними критеріями. Динамічний розвиток корпусної лінгвістики за доволі недовгий час призвів до виділення трьох основних проблем і, відповідно, внутрішніх напрямів в межах корпусного мовознавства, а саме: 1) теорія і практика побудови корпусів текстів природних мов, 2) теорія і практика програмного оброблення корпусних даних та 3) теорія і практика застосування корпусів як у лінгвістичних, так і інших типах досліджень, які звертаються до даних природної мови.

Корпусна лінгвістика первинно сформувалася на ґрунті англо-саксонського мовознавства, тому не випадково загальний огляд англomовної бібліографії з найрізноманітніших проблем корпусного мовознавства сягає понад тисячу позицій. Особливе місце у ній займають праці Х. Кучери [1], У. Френсиса [2], Дж. Синклера [3], В. Тойберта [4], Г. Кеннеді [5] etc. Хоча й не дуже активно, але проблеми корпусного мовознавства висвітлюють російські та білоруські мовознавці. Цікавими серед публікацій останніх є праці А. Баранова [6], С. Шарова [7], В. Рикова [8], Л. Ричкової [9].

Актуальність нашої статті зумовлена тим, що українська лінгвістична традиція на сьогодні не може похвалитися наявністю у своїй науковій парадигмі самостійного напрямку, яким є корпусна лінгвістика в англо-саксоністиці, романо-германістиці та частині славістики. Усі дослідження тут зведено до спроб започаткувати цей напрям (див. публікації „Корпус текстів української періодики” [10], „Базові поняття корпусної лінгвістики” [11]). Але стверджувати, що спроби започаткувати повноцінний напрям корпусного мовознавства в україністиці відбуваються на, так би мовити, порожньому місці, немає

підстав. Можна вважати підґрунтям цих досліджень, зокрема, праці Н. Дарчук [12], Є. Карпіловської [13], Н. Клименко [14], В. Перебийніс [15], М. Пещак [16] та інших українських мовознавців.

Побудова довільного корпусу неможлива без створення попереднього деталізованого проекту, у якому, крім параметризації предметної галузі, яку репрезентуватиме корпус, визначення загальнокорпусного обсягу і статистичних параметрів його конститутивних текстів, засад кодування структури корпусу й анотування лінгвальних корпусних даних etc., обов'язково необхідно визначити релевантний набір і структуру метаданих проєктованого корпусу, де під метаданими розуміємо **набір екстралінгвістичної інформації про текстовий матеріал корпусу загальномовного типу**.

Поліаплікабельність загальномовного корпусу передбачає для себе якісно різного користувача – лінгвіста vs нелінгвіста. Якщо користувачем є лінгвіст, то слід передбачити, що лінгвісти різних спеціалізацій ставитимуть різні вимоги як до самого корпусу, так і до мовних даних, отриманих із корпусу. Як зазначає А. Баранов, для досліджень із морфології та синтаксису зазвичай не потрібні надто великі корпуси, а інколи навіть небажано звертатися до них за даними, оскільки, наприклад, число цитат „вживання службових слів на взірець *або, так, ні* може досягнути у гіперкорпусі кількох тисяч сторінок. На один цікавий приклад може припадати сотня тривіальних” [17, 119]. Тому для дослідження граматики важливою є не кількість, а якість текстового матеріалу, який повинен варіюватися структурно і жанрово. Але в загальномовному корпусі усе ж не слід звужувати дані до мінімуму, тому що мінімалізація може стати причиною неохоплення рідковживаних слів чи специфічних синтаксичних конструкцій, а отже передумовою невиявлення певних граматичних ознак мови. Натомість користувач-нелінгвіст залишить поза увагою лінгвальний аспект і його, найімовірніше, зацікавить екстралінгвальна інформація, наприклад, дата, місце, передумови створення тексту, вік і стать автора тексту, кількість сторінок, видавництво, принципи кодування й анотування тощо, тобто **метадані**.

Виходячи із загальних принципів проєктування корпусів текстів, їх автори часто пропонують власний набір метаданих. На нашу думку, одним із найкращих прикладів індивідуальної детермінації набору і структури корпусних метаданих є система метаданих у *Динамічному корпусі сучасної російської публіцистики 90-х років*. Вимога індивідуального набору параметрів була зумовлена потребою інвентаризувати проблемну галузь, репрезентовану цим корпусом. Так, система метаданих *Динамічного корпусу сучасної російської публіцистики* детермінована набором факторів, принципово важливих, з погляду його авторів, для забезпечення інформативності корпусних даних і йдеться про:

- фактор автора тексту: журналіст / непрофесійний політик vs професійний політик (до уваги взято як політичних діячів першого порядку, так і другого); окремо стоїть проблема виявлення команд спічрайтерів, які визначають власне мовне оформлення тексту;
- фактор персоніфікації-деперсоніфікації автора (конкретна людина vs партія / суспільний рух / політична організація / установа vs деперсоніфікований текст – лозунги, передовиці тощо);
- фактор адресата (кому адресований текст: прибічники – противники – нейтральна аудиторія; професійна орієнтація – виступ перед шахтарями; творчою інтелігенцією тощо);
- фактор прагматичних умов генерації тексту (промова на мітингу – промова на засіданні інституційного органу – інтернет-, прес-конференція);
- фактор джерела: журнальний текст – книжковий текст – листівка – агітаційний плакат – лозунг – телебачення – радіо;
- комунікативний розподіл (монологічний текст – діалог; загальні типи ілюцій: демонстрування намірів, наприклад, політична програма – аргументацій діалог тощо).

На базі наведених факторів була укладена матриця параметрів, яка дозволила виділити з предметної галузі приблизно 70 типів текстів, а далі сформульована типологія лягла в основу відбору текстів до корпусу.

Дещо по-іншому формує систему параметрів і, відповідно, набір метаданих В. Андрющенко [18, 15-16]:

- I. Стандартний бібліографічний опис, який складається з:
 1. Імені автора з ініціалами після прізвища;
 2. Повної назви та ідентифікації джерела.
- II. Копірайт.
- III. Ідентифікатор комп'ютерної версії тексту.
- IV. Ідентифікатор жанру: проза, поезія, драма, преса, науковий текст etc.
- V. Ідентифікатор системи кодування.
- VI. Ідентифікатор предметної області для наукових і науково-популярних жанрів.
- VII. Ідентифікатор тому.
- VIII. Ідентифікатор частини.
- IX. Ідентифікатор книги.
- X. Ідентифікатор розділу.
- XI. Ідентифікатор глави.
- XII. Ідентифікатор дії.

Цей набір відповідає стандартному наборові, сформульованому в корпусному мовознавстві та базованому на схемі опису документа TEI, яка складається з двох обов'язкових частин: 1) опису тексту (передбачає наявність даних про автора твору, заголовок, назву видання, видавництво, місце видання, дату, кількість сторінок) і 2) опису профілю електронного документа (передбачає наявність інформації про стиль і

жанр тексту, загальну тематику, додаткових даних про автора: вік, стать, місце постійного проживання, цільову аудиторію, наявність / відсутність перевидань тексту, розмір тексту в словах і байтах, назву файлу, у якому збережено електронний варіант тексту, набір тегів для розмітки структури документа і морфологічної анотації, обставини створення тексту тощо). Перша частина фактично відповідає традиційній бібліографічній інформації, яку зазвичай подають у лексичній картотечній картці, у посиланнях та списку літератури. А друга частина забезпечує два типи інформації – прагматичну та технічну. І залежно від дефініції корпусу, ці дві групи інформації можуть подаватися повністю або частково. Але, як правило, набір параметрів опису профілю документа повністю не забезпечують.

Будуючи український корпус загальномовного типу, доцільно використати стандартний TEI-підхід до метаданих, тобто усі метадані згрупувати як: а) дані, що описують первинний текст і б) дані, що описують профіль електронного тексту.

Опис первинного тексту в корпусі загальномовного типу передбачає набір таких параметрів:

- автор / автори тексту;
- заголовок тексту;
- назва видання для текстів, опублікованих у збірниках, часописах, колективних монографіях тощо;
- том і / або серія, і / або випуск, і / або номер, і / або частина для багатотомних та періодичних видань;
- назва тому, випуску (якщо є);
- місце видання;
- назва видавництва;
- рік видання;
- кількість сторінок.

І відповідно, кожен з визначених параметрів повинен мати власне значення, записане у стандартному форматі:

- автор: для української традиції притаманні дві моделі подання власних назв людей, що призводить до варіантності формалізації значення до параметра „автор”: а) двокомпонентна модель передбачає запис „прізвище + ім'я”; б) трикомпонентна – „прізвище + ім'я + по батькові”, але обидві моделі вимагають повного запису своїх елементів, наприклад: *Франко Іван, Українка Леся, Ющенко Віктор, Кобилянська Ольга, Шевченко Тарас Григорович, Чепіга Іна Петрівна, Коломієць Сидір Іванович*. Не залежно від прийнятої моделі подання антропоніма, нормативним для корпусу записом (а корпусна норма, як правило, відповідає загальномовній нормі) є „прізвище + ім'я + по батькові” повністю і цей запис, якщо його відразу не реалізовано, досягають простим програмним правилом перестановки компонентів антропоніма. Така стратегія зумовлена вимогою стрункості організації корпусних даних і спрощення пошуку в межах корпусу.

○ заголовок: значенням до цього параметра є класичний текстовий заголовок, наприклад, *Заповіт*, *Борислав сміється*, *Структура іменувань у програмуванні*, *Українська мова в чеському мовознавстві*;

○ назва видання: значенням є назва джерела, у якому опубліковано текст, наприклад, *Наукові записки НаУКМА* (збірник), *Українська мова* (часопис); *Урядовий кур'єр* (газета);

○ том, серія, випуск, номер: значення параметра „том” слід подавати і до багатотомного видання творів, і тоді, коли періодичне видання нумеровано поточно; серія та випуск головно задають стосовно періодичних видань і, як правило, наявність значення за параметром „серія” виключає значення за параметром „випуск”, „номер”, „частина”, і навпаки, але якщо існують два і / або більше параметри, то їх значеннями є відповідні цифрові дані, наприклад, ... *Наукові записки НаУКМА*. – Т. 16. [...] №1 ...; ... *Зібрання творів Франка*. – Т. 25 ...; ... *Українська мова*. – №3 ..., ... *Урядовий кур'єр*. – Вип. 224;

○ назва тому, випуску: значенням є назва джерела, у якому опубліковано авторський текст, наприклад, *Наукові записки НаУКМА*. – Т. 16. *Комп'ютерні науки* ...; ... *Науковий вісник Львівського університету*. – Вип. 21. *Філологічні науки* ...;

○ місце видання: значенням є географічна назва населеного пункту, у якому видано описуване джерело, записана повністю, наприклад, ... *Київ* ..., ... *Львів* ..., ... *Париж–Сарсель* ..., *Москва* ...

Забезпечуючи значенням параметр місця видання у системі корпусних метаданих, не можна використовувати звичний для української бібліографічної традиції запис М. – Москва, К. – Київ, Л. – Ленінград, оскільки це створює передумови для машинної неоднозначності, наприклад Л – Львів, Ленінград, Луцьк etc., а отже програмної помилки, і вимагатиме додаткових формальних правил для програми перетворення такого запису.

○ назва видавництва: значення як правило реальне для книжкових текстових джерел класу монографій, зібрань творів, словників, наприклад: ... *Наукова думка* ..., ... *Пульсари* ..., ... *Основи* ...;

○ рік видання: значенням є рік, записаний цифрами, наприклад: *Пульсари*, 2002, ... *Наукові записки НаУКМА*. – Т. 16. *Комп'ютерні науки*, 1999;

○ кількість сторінок: значенням є цифровий запис загальної кількості сторінок у першоджерельному тексті для книжкового або словникового видання і номер першої та останньої сторінки тексту в збірниках, записаний через тире, не зважаючи на те, чи до корпусу текст увійшов повністю, чи лише його фрагмент, наприклад, 234, 24–54.

Другу частину метаданих українського корпусу формують такі параметри:

- жанрово-стилістична ідентифікація тексту;
- територіальна ідентифікація тексту;
- інформація про видання тексту;

- вікова і статеві інформація про автора тексту.

Аналогічно кожному з цих параметрів повинні відповідати стандартні значення:

- жанрово-стилістична ідентифікація тексту: набір релевантних значень детермінований жанрово-стилістичною структурою конкретного корпусу.

Наприклад, у проекті Українського національного корпусу йдеться про максимально семикомпонентну стилістичну диференціацію української мови і, відповідно, задовільним значенням стилістичної ідентифікації тексту буде 1) художній, 2) науковий, 3) офіційно-діловий, 4) публіцистичний, 5) конфесійний, 6) епістолярний і 7) розмовний стилі; а значенням жанрової ідентифікації, наприклад, для художнього стилю буде: а) проза, б) поезія, в) драматургія.

- територіальна ідентифікація тексту: задовільне значення – вказівка на, по-перше, місце створення тексту: *Сх. Україна – Харків, США – Ванкувер* (коли йдеться про діаспорні тексти); і, по-друге, місце сталого проживання автора тексту: *Зах. Україна – Дрогобич, Словаччина – Свидник*; для діалектного матеріалу додатково слід забезпечити ареальну ідентифікацію тексту;

- інформація про перевидання тексту: значення цього параметра складається з трьох елементів 1) номер видання / перевидання, 2) примітка про тип перевидання і 3) рік відповідного видання / перевидання, наприклад: *1 – 1623; 2 – доповнене – 2001, 3 – препринт – 1987;*

- вікова і статеві інформація про автора тексту: значення цього параметра, передовсім актуальне для соціолінгвістичних досліджень, формується з дати народження, записаної у форматі „день + місяць + рік”, і буквені позначки статі, яка відповідає нормативному позначенню родів у словниках і граматиках української мови – *ж., ч.*

Отже, вимога стандартності побудови сучасних корпусів текстів загальномовного типу та їх поліаплікабельність передбачає не лише встановлення релевантних загальних вимог до корпусу, здійснення попереднього детального планування його побудови, окреслення текстової джерельної бази, але й обов'язковість формалізації та подання метайнформації про конститутивні корпусні тексти, що перетворює довільне зібрання текстів у електронній формі на корпусну побудову й уможливорює застосування корпусу широким колом користувачів як лінгвістів, так і соціологів, істориків, програмістів etc.

Література

1. **Francis W. N., Kucera H.** A Standard Corpus of Present-Day Edited American English (Brown corpus). – Providence, 1979.
2. **Френсис У.** Проблемы формирования и машинного представления большого корпуса текстов // Новое в лингвистике. – 1983. – Вып. XIV. – С. 334-353.
3. **Sinclair J.** Corpus Typology Draft. – <http://www.icl.pi.cnr.it>, 1994.

4. Teubert W. Corpus Linguistics – a Partisan View // International Journal of Corpus Linguistics. – 2000. – Vol. 5. – No.1. – P. 3-27.
5. Kennedy G. Introduction to Corpus Linguistics. – London-New-York, 1998.
6. Баранов А. Н. Введение в прикладную лингвистику. – М., 2001.
7. Шаров С. А. Большой Корпус русского языка. – www.bokrcorpora.narod.ru, 2002.
8. Рыков В. Корпусная лингвистика. – <http://rykov-cl.narod.ru/lekci.doc>, 2001.
9. Рычкова Л. В. Корпусная лингвистика: лексикографический аспект // Слово и словарь. – Гродно, 2002. – С. 182-189.
10. Демская-Кульчицкая О. М. Корпус текстов украинской периодики // Исследование славянских языков в русле традиций сравнительно-исторического и сопоставительного языкознания: Информационные матер. и тезисы докладов междунар. конф. – М., 2001. – С. 26-28.
11. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики // Укр. мова. – 2003 – №1.
12. Дарчук Н. П. Частотний словник сучасної поетичної української мови. – <http://www.philolog.univ.kiev.ua>, 2000.
13. Карпіловська Є. Н. Вступ до комп'ютерної лінгвістики. – Донецьк, 2003.
14. Клименко Н. Ф. Построение тезауруса с помощью ЭВМ // Украинский семантический словарь. Проспект. – К., 1990.
15. Перебийніс В. С. Теоретичні та прикладні проблеми структурно-математичної лінгвістики // Мовознавство. – 1981. – № 4. – С. 28-34.
16. Пещак М. М. Нариси з комп'ютерної лінгвістики. – Ужгород, 1999.
17. Баранов А. Н. Там само.
18. Андрищенко В. М. Об организации архива источников Машинного фонда русского языка, их разметке и комментировании // Бюллетень Машинного фонда русского языка. – М., 1992. – Вып. 1. – С. 14-19.

The article deals with the problem of metadata in general Ukrainian texts corpus, especially its structure, main parameters of the metadata and standard sense of these parameters and stressed on importance of formalization of the metainformation for corpora.

УДК 81'1

А. С. Зеленько

ВЗАЄМОДІЯ МЕНТАЛЬНО-РАЦІОНАЛЬНОГО Й ПОЧУТТЄВО-ЕМОЦІЙНОГО У ФУНКЦІОНУВАННІ Й СТРУКТУРІ МОВИ

Ця розвідка розв'язує часткове питання раціонального й почуттєвого у становленні значення – у широкому його розумінні – у мові та структурі основних одиниць мовної системи, що постає як один з основних принципів лінгвістичного детермінізму у витлумаченні її автора. Це і ряд інших розвідок, підготовлених і надісланих на різні наукові форуми, найближчим часом складуть ще одну книгу, в якій завершується оформлення лінгвістичного детермінізму під кутом зору синтезу здобутків радянського мовознавства та когнітивної лінгвістики.