УДК 81’33

Монахова Т.В.

ЗАСТОСУВАННЯ ПРИЙОМІВ КОРПУСНОЇ ЛІНГВІСТИКИ В ЛЕКСИКОГРАФІЇ

У статті розглянуто можливість ефективного застосування корпусного підходу до укладання словника мови автора. Розкрито основні проблеми створення авторського корпусу як підґрунтя для укладання ідіолектного словника. Подано зразок ідеологічного словника у вигляді тезаурусу.

Ключові слова: *ідіолект, корпусна лінгвістика, лексикографія.*

В статье рассматривается возможность эффективного применения корпусного подхода к составлению словаря языка автора. Раскрыты основные проблемы создания авторского корпуса как основания для составления идиолектного словаря. Приведен образец идиологического словаря в виде тезауруса.

Ключевые слова: *идиолект, корпусная лингвистика, лексикография.*

In this article the possibility of effective use of corpus-based approach to the compiling of an author’s language dictionary is considered. It analyzes the main problems of drawing up an author’s corpus as a basis for compiling an idiolect dictionary. We propose an example of ideological dictionary as a thesaurus.

Key words: *idiolect, corpus-based linguistics, lexicography.*

Дослідники мови все частіше говорять про вихід мовознавства з кола суто гуманітарних наукових дисциплін і надання йому якостей дисципліни технологічної, стверджуючи, що в постіндустріальному суспільстві природна людська мова – мабуть, уперше в історії людської цивілізації – набуває технологічного статусу, від якого починає безпосередньо залежати ефективність функціонування суспільного виробництва, а зрештою – й інших суспільних інститутів. Лінгвістичний корпус стає підґрунтям створення інформаційно-лінгвістичних артефактів, призначених для інтелектуального опрацювання мови й необхідних для створення й функціонування високоефективних технологій оперування знаннями.

Проблематика корпусного напряму доволі розгалужена і передбачає студії, по-перше, загальної теорії корпусної лінгвістики, над якою працюють, зокрема, Дж. Синклер [12], В. Тойберт [13]; по-друге, кореляції корпусної лінгвістики та інших лінгвістичних дисциплін, по-третє, типології корпусів і методики інтерпретації корпусних даних, по-четверте, засад створення текстових корпусів природних мов тощо. Йдеться про доробок М. Баньки [10], Ґ. Кеннеді [11], О. Демської-Кульчицької [3], В. Рикова [5], Л. Ричкової [6], С. Шарова [9], В. Широкова [4] та ін.

Розробленням методики та процедури організації лексикографічного матеріалу на засадах корпусної лінгвістики займаються науковці Національного корпусу відділу лексикології, лексикографії та Національного корпусу української мови Інституту української мови НАН України, зокрема, це створення електронного варіанта «Словаря української мови» за ред. Б. Грінченка [2].

Новизна нашої роботи полягає у спробі застосування дослідницьких корпусних методик до укладання словника мови автора, що мотивоване передовсім появою та інтенсивним розвитком такого напряму в сучасній науці про мову, як корпусна лінгвістика, яка уможливлює вивчення, аналіз, інтерпретацію тощо мовних реалій дещо під іншим кутом зору, ніж у так званій класичній, традиційній лінгвістиці. Погляд під іншим кутом зору на мовні реалії, відповідно, дає змогу отримувати нові результати й робити нові висновки. Швидкий розвиток інформаційних технологій, розширення електронних бібліотек та зростання кількості літератури на електронних носіях, зокрема й словників, зумовили актуальність нашого дослідження. Потреба у створенні ідіолектних словників класиків української літератури є очевидною.

Історично першою спробою використання комп’ютера для лінгвістичних цілей вважають Центр автоматизації вивчення літературних текстів у Галараті (Італія, 1956 р.). Основним завданням Центру було введення до комп’ютера максимального числа текстів, причому в різній графіці, з метою генеральної інвентаризації різноманітних лінгвістичних фактів, а саме: створення різного роду покажчиків і конкордансів слів, морфем, графем,

синтаксем, частот, тобто матеріалу, який у подальшому планувалося використовувати для лінгвістичних, психологічних ті інших досліджень.

У сучасній корпусній лінгвістиці, крім побудови загальномовних корпусів, поширена практика створення й інших корпусів, серед яких значне місце посідають *авторські корпуси*.

Авторськими називають повнотекстові корпуси одного тексту, наприклад корпус Біблії, або текстів одного автора. Популярним авторським корпусом є *Корпус текстов словаря языка Достоевского*, який створювався як джерело для словника мови Ф. Достоєвського, і на сьогодні вже існує його CD-версія під назвою *«Достоевский: Тексты, исследования, материалы»*, яку поширюють разом із програмами оброблення корпусу, базою даних із ідіоматики і базою даних частотного словника. Укладено «Словарь языка Достоевского. Лексический строй идиолекта», який складається з кількох випусків і будується на словнику найважливіших для автора лексичних одиниць, які презентують світ мовної особистості [7].

Створено *Корпус мови Пушкіна*, за матеріалами якого укладено «Словарь языка Пушкина» [8]. Це найбільш повний і теоретично розроблений словник мови письменника. У ньому описано понад 20 000 слів російської мови, що зустрічаються у художніх і публіцистичних творах О.С. Пушкіна, а також у його листах і ділових паперах. Для кожного слова подано словникову статтю, в якій показано кількість його слововживань у текстах Пушкіна, сформульовано його значення, проілюстровано цитатами й запропоновано повний перелік слововживань, що містять указівки на граматичні форми й посилання на всі тексти, в яких зустрічається це слово. Окремо репрезентовано функціонування слова у складі фразеологічних одиниць.

Так, обов'язковими корпусними ознаками, згідно з твердженням О. Демської-Кульчицької [3, с. 56-58], є:

1. **Відібраність,** яка ставить вимогу відбору фактичного матеріалу. Виникає необхідність створення певної вибірки, яка передбачає застосування чітких правил екстрагування даних, що відповідають обраній стратегії побудови корпусу, мотивовані типом корпусу і метою його створення та специфікою експлуатації.
2. **Репрезентативність,** яка полягає в здатності корпусу відображати всі властивості предметної галузі.
3. **Збалансованість,** що полягає у введенні до корпусу пропорційної кількості текстових ресурсів. На практиці, де традиційно використовують різні методики відбору текстового матеріалу до корпусу, одним із доволі складних завдань є досягнення збалансованості. Для досягнення збалансованості корпусу необхідні мінімальні критерії відбору текстів, які мають включати розрізнення між художньою літературою і нехудожньою літературою; книжкою, журналом або газетою; нормативним і ненормативним варіантом мови; з контролем віку, статі та походження авторів.
4. **Машиночитаність** є визначальною ознакою до сучасного електронного текстового корпусу природної мови. Крім електронної форми подання, ця вимога передбачає наявність кодування первинних корпусних даних та лінгвістичну анотацію, хоча на сьогодні це вже параметр «за промовчанням», тобто іншим сучасний корпус не повинен бути.
5. **Стандартність** забезпечує узаконене, однозначне, мовнонезалежне оброблення даних довільної природної мови. Категорія стандартності в дослідній парадигмі корпусної лінгвістики функціонує паралельно з категорією нормативності, але, на відміну від останньої, має чіткий технічний характер, що уможливлює аналогічні дослідження на багатьох корпусах, еволюцію самого корпусу і як безпосереднє синхронне, так і діахронне використання корпусного ядра даних.

Формат розмітки становить чи не основну проблему корпусного анотування. Важливим у корпусній лінгвістиці залишається питання створення засобів автоматичного (чи, принаймні, автоматизованого) анотування текстів за різними критеріями – морфологічними, орфоепічними, семантичними, синтаксичними тощо [4, с. 7]. За В.А. Широковим, основна ідея системотехніки лінгвістичного корпусу (випускаючи технічні деталі) полягає у забезпеченні автоматичного розбиття електронного тексту літературного джерела на «мікроконтексти» – фрагменти тексту, які «групуються» навколо слова, що є об'єктом тлумачення. Таким чином, відпадає необхідність формування і збереження традиційного для мовознавства об'єкта – лексичної картки як окремого фізичного об'єкта – вона перетворюється на об'єкт віртуальний, тобто реалізований як певне відношення у відповідній базі даних. «Справді, при цьому підході достатньо забезпечити виконання процедури природномовної індексації тексту, що означає приписування кожній текстовій словоформі її формально визначеної локалізації – своєрідної координати в розглядуваному тексті з наступною лематизацією цієї словоформи» [4, с. 99].

Аплікативне призначення корпусних даних – фонологічні, морфологічні, синтаксичні, лексикографічні, лексикологічні тощо дослідження – детермінує тип лінгвістичної анотації корпусу. Як правило, фонетична анотація має формат фонетичної транскрипції. До морфолого-синтаксичної (чи морфо-синтаксичної, за терміном корпусної лінгвістики) анотації існує два підходи: перший передбачає синтез морфологічного та синтаксичного аспектів анотування, йдеться про граматичну анотацію; другий підхід передбачає диференціацію морфологічної та синтаксичної анотації. Під морфологічною анотацією розуміємо тип лінгвістичної анотації, за якої експліковано подається

морфологічна інформації про тексові елементи рівня слова [3, с. 112].

Традиційно в корпусній лінгвістиці під лінгвістичною анотацією розуміють:

- довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних;
- практику введення формалізованої лінгвістичної інформації в електронний текст;
- наявність такої інформації в тексті.

Схильна до розрізнення термінів *анотація й анотування* О. Демська-Кульчицька [3]. Так дослідниця зазначає, що *анотація* – ще певна лінгвістична інформація про лінгвально релевантні одиниці текстових даних і наявність такої інформації у тексті; *анотування* – це процес уведення формалізованої лінгвальної та лінгвістичної інформації в електронний текст.

Реалізація будь-якого типу анотування передбачає низку процедур:

1. Сегментизація тексту – ідентифікація та виділення концептів і їхніх іменникових репрезентантів.
2. Формалізація параметрів анотування цільових одиниць маркування.
3. Створення тегсету, чи набору формальних кодів з відповідною семантикою, засобами яких адекватно детерміновано для кожної цільової одиниці тексту її відношення до повного опису ідеолекту письменника.
4. Визначення анотаційної схеми та її принципів.

Автори монографії «Корпусна лінгвістика» [4] говорять про такі критерії застосування стандарту: 1) *достатність* – набір структурних елементів повинен бути достатньо широким, щоб забезпечити хоча б більшість вимог. Водночас бажано, щоб схема розмітки не містила надлишкову інформацію; 2) *несуперечливість* – схема розмітки має бути сформована на базі несуперечливих правил, які б дозволяли однозначно визначити, які об'єкти належать до тегів, які – до атрибутів, що є вмістом тега тощо; 3) *відтворюваність* – схема кодування повинна ґрунтуватися на чітко визначених правилах, що дає можливість відтворення вихідного тексту за допомогою простих алгоритмів; 4) *коректність* – за допомогою спеціального програмного забезпечення відбувається перевірка відповідності міток у документах їх структурним специфікаціям; 5) *можливість збору даних* – збір даних включає безпосереднє накопичення даних (за допомогою ручного вводу або з використанням автоматичного розпізнання тексту) та проведенням кодування даних; 6) *технологічність* – урахування потреб, пов'язаних з автоматичною обробкою текстів (вибір тексту згідно зі встановленими критеріями, використання спеціальних механізмів, типу міжтекстових покажчиків, поєднання текстів або інших елементів корпусу) тощо; 7) *можливість масштабування* – важливо, щоб будь-яка створена схема мала можливість поповнюватися; 8) *компактність* – проведення розмітки може істотно вплинути на розмір файлу, від чого залежить швидкість обробки даних текстів. Серед можливих методів досягнення компактності називають: мінімізацію тегу, наприклад, пропущення або скорочення кінцевого тегу; застосування специфічних кінцевих тегів елементів або відмова від останніх; використання XML схеми розмітки тощо; 9) *зрозумілість* – коли виникає потреба у безпосередній роботі користувача з текстом без використання спеціального програмного супроводу, прозорість розмітки є досить важливою [3, с. 51-53].

З погляду типологічно-аплікативних характеристик авторський корпус можемо розглядати як:

- *ілюстративний*: створюватиметься після визначення й детального вивчення ідіолекту письменника;
- *повнотекстовий*: збудований із цілих текстів творів автора;
- *статичний*: не передбачає перманентного поповнення множини корпусних текстів;
- *синхронний*: охоплює рівень сучасної української мови;
- *мови автора*: тільки тексти одного письменника входять до складу корпусу;
- *мономовний*: усі тексти є результатом мовної діяльності носія-кодифікатора сучасної української мови;
- *писемний*: корпус становить зібрання писемних текстів;
- *концептно анотований*: текстові дані предметної галузі розмічені до рівня фраземи з частковим граматичним маркуванням іменникових репрезентантів.

На основі створеного авторського корпусу маємо змогу укласти тезаурус. **Тезаурус** – це словник, що відображає весь словниковий склад мови з вичерпним переліком прикладів вживання слів у тексті. Це ідеографічний словник, а також інформаційно-пошуковий довідник, в якому перелічені всі лексичні одиниці інформаційно-пошукової мови (дескриптори), усі синонімічні їм слова, словосполучення й регулярні семантичні відношення між дескрипторами.

Оскільки ідіолект письменника становить не тільки «сукупність індивідуальних (професійних, соціальних, територіальних, психофізичних та ін.) особливостей, що характеризують мовлення певного індивіда; індивідуальний різновид мови» [1, с. 165], а й відображає сукупні менталітетні риси певної національної мови поряд із авторськими, індивідуальними рисами, то тезаурус мови письменника може мати вигляд ідеологічного словника.

Ідеологічний словник – словник, у якому слова розташовані у вигляді тематичних рядів [1, с. 421]. Такими тематичними рядами можуть слугувати, наприклад, концепти, що утворюють концептосферу ідіолекту певного письменника. Концептосфера об'єднує всі семантичні поля, які номінують і характеризують найважливіші поняття. Сама будова кожного концепту є рівневою: складається з ядра, приядерної зони та периферії концепту. Тезаурус концепту може мати такий вигляд:

Тезаурус концепту *дім* в ідіолекті Валерія Шевчука

1. Ядро концепту *дім*

– будинок
– будівля
– господа
– дім
– домівка
– житло
– квартира
– кімната
– маєток
– мешкання
– оселя
– осілля
– обитель
– обійстя
– подвір'я
– помешкання
– родина
– рід
– садиба
– споруда
– хата

2. Приядерна зона концепту *дім*

2.1. Складники *дому*

– бляха
– вітальня
– віталенька
– вікна
– віконечка
– віконниці
– віконця
– ворота
– горище
– ґанок
– долівка
– зала
– дах
– двері
– двір
– каркас
– кватирки
– коридор
– коридорець
– коридорчик
– кухня
– одвірок
– сінці
– стіни
– стріха
– сходи
– паркан
– підвал
– підвіконня
– під'їзд
– підлога
– передухіддя
– поверх
– поріг
– тин
– фіранки
– центральний вхід
– чорний вхід
– шиби
– хвіртка

2.2. Різновиди *дому*

2.2.1. «*Дім другої половини XX століття*»

– дев'ятиярусна коробка
– кімнатки
– кімната в комуналці
– мікроквартира
– міська клітка
– новочасні коробки
– однокімнатка
– п'ятиповерхова коробка стилю «бароко»
– сучасна коробка
– трикімнатна квартира
– трикімнатка

2.2.2. «Тимчасовий *дім*»

– будинок-флігель
– дача
– дачка
– літній домик
– літня кухня
– сарай
– халабуда

2.2.3. «Великий *дім*»

– замок
– замочок
– замчисько
– опочивальня
– палати
– палац
– палац стилю «бароко»
– передпокій
– передпокійчик
– покій
– покоєць
– покоївок
– покоїк
– світлиця
– спочивальня
– фортеця
– хороми

2.2.4. «Малий *дім*»

– будиночок
– закапелок
– келія
– келія-печера
– кімнатка
– комірчина
– куща
– хатка
– хижа

2.3. Антропоморфні назви, пов'язані з концептом *дім*

– господар
– домашні
– домовик
– домочадці
– квартирант
– келійник
– мешканці

– пожильці дому
– поселенець
– посельці
– співжильці

3. Периферія концепту *дім*

3.1. Дім як будівля для розміщення різних закладів і установ

– будинок відпочинку
– будинок народної творчості
– будинок офіцерів
– будинок творчості письменників
– гуртожиток
– дім-вбиральня
– дім-хроніка
– дитбудинок
– дитячий будинок
– дурдом
– корчемний дім
– мисливська хатина
– мисливська хатка
– молебний дім
– молитовний дім
– явочні квартири

3.2. Фразеологічні одиниці, що увиразнюють концепт *дім*

– батьківський дім
– біду в дім узяти
– виносити сміття з хати
– іграшковий дім
– мати власне гніздо
– моя хата скраю
– на поріг не пускати
– ні кола, ні двора
– оббивати пороги
– переступити чийсь поріг
– під дурного хату
– прийняти в дім
– по чужих кутках
– показати комусь на двері
– рідна оселя
– розриватися між домом і роботою

3.3. Переносні значення з експліцитно вираженою позитивною оцінкою

– барліг
– гніздо
– гніздо родинне
– гавань
– куток
– місце для сталого прожитку
– парубоцька нора
– сімейна спілка
– тихе царство
– цегляний велетень

3.4. Переносні значення з імплікованим (прихованим) емоційним забарвленням

– башта слонової кості
– капсула
– хатини-вози
– острів-замок
– стіна-екран
– хати-криївки
– «людина нори»

3.5. Переносні значення з експліцитно вираженою негативною оцінкою

– Залізний Мішок
– комірчина
– конура
– купа цегли, дощок, балок
– мішок
– пекло
– руїни
– сірі мури
– сірі стіни
– стерильна порожнеча
– тісне та скорбне місце

3.6. Юридичний процес набуття людиною власного житла

– перепрописка
– прописка

Як бачимо, тезаурусна презентація одного концепту концептосфери письменника, зокрема концепту «дім» в ідіолекті Валерія Шевчука, вже охоплює значний обсяг лексичних одиниць, які супроводжуються великою кількістю прикладів, або контекстних уживань, кожної лексеми, адже в електронному словникові кожне слово є активним, тобто при наведенні курсора розкривається вікно з усіма можливими прикладами його вживання у текстах письменника. Так розвиток сучасної прикладної лексикографії передбачає забезпечення максимальної інформативності лексикографічних систем.

Застосування корпусного методу до організації та укладання ідіолектних словників відкриває великі перспективи перед лінгвістами. Предметом дослідження можуть бути не тільки лексичні значення (концептосфери), а й морфологічні, синтаксичні й навіть фонетичні особливості мови кожного автора.

ЛІТЕРАТУРА

1. Ахманова О.С. Словарь лингвистических терминов. – М.: КомКнига, 2007. – 576 с.
2. Балог В., Балог О. «Словарь української мови» за ред. Б. Грінченка (1907-1909 рр.): електронна версія // Лексикографічний бюлетень: Зб. наук. пр. – К., 2007. – Вип. 16. – С. 78-82.
3. Демська-Кульчицька О.М. Основи національного корпусу української мови. – К.: Наукове видання Ін-ту укр. мови НАН України, 2005. – 219 с.
4. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с. – Бібліогр.: С. 459-467.
5. Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы. – Режим доступу: http://rykov-cl.narod.ru, 2001.

6. Рычкова Л.В. Корпус как совокупность многослойных полнотекстовых баз данных // Труды международной конференции «Корпусная лингвистика – 2004». – СПб.: Изд-во С.-Петерб. ун-та, 2004. – С. 316-323.
7. Словарь языка Достоевского. Лексический строй идиолекта / Российская академия наук. Институт русского языка им. В.В. Виноградова; Гл. ред. Ю.Н. Караулов – М.: Азбуковник, 2001.
8. Словарь языка Пушкина: В 4 т. / Отв. ред. В.В. Виноградов. – 2-е изд., доп. / Российская академия наук. Ин-т рус. яз. им. В.В. Виноградова. – М.: Азбуковник, 2000. – 982 с.
9. Шаров С.А. Большой корпус русского языка. – Режим доступу: www.bokrcorpora.narod.ru, 2002.
10. Banko M. Korpus tekstow jako zrodlo wiezdy o jezyku // Wyklad na sesji MSH Universytetu Warszawskiego: Rekopis. – Warszawa: PWN, 2003.
11. Kennedy G. Collocatios: Where grammar and vocabulary teaching meet // Language teaching methodology for the nineties. – Singapore: RELC Anthology. – 1990. – Series 24. – P. 215-229.
12. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991. – 137 p.
13. Teubert W. Corpus Linguistics and Lexicography // International Journal of Corpus Linguistics – 2001. – Vol. 6. – Special issue. – P. 125-153.

Рецензенти: д.філол.н., професор Бєлєхова Л.І.,
д.філол.н., професор Корольова А.В.

Надійшла до редколегії 19.11.2008 р.