I. МОДЕЛЮВАННЯ ЛІНГВАЛЬНИХ ЯВИЩ І НОВІТНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Текстозорієнтована ІПС з літературознавчої термінології

Наталія Дарчук

д. філол. н., професор, професор кафедри української мови та прикладної лінгвістики, Інститут філології Київського національного університету імені Тараса Шевченка, Україна, E-mail: nataliadarchuk@gmail.com

The article describes lexicographical and encyclopedic electronic database of Ukrainian literary terms, which consists of three dictionaries: alphabetical, explanatory and thesaurus of 1350 terms. The technology of data retrieval thesaurus system is created. It may function as a reference and encyclopedic system or inside the other IS.

Ключові слова — термін, інформаційно-пошукова система, тезаурус, тезаурусні відношення.

Вступ

Для забезпечення високого рівня проведення конференції рекомендуємо авторам дотримуватися вимог, поданих у зразку.

Оргкомітет рекомендує подавати статті в MS Word 200x, використовуючи цей шаблон.

Будь-які роботи в галузі автоматичного опрацювання тексту (автоматичний переклад, автоматичне реферування й анотування тексту, індексування тощо) зупиняються перед семантичним бар'єром. Відсутність повних описів семантики природних мов не дозволяє сподіватися на швидке здолання цього бар'єру. Прийшов час довготривалої роботи з укладання таких описів, пошуку, тестування на широкому корпусному матеріалі різних методів аналізу і синтезу мовних значень.

Одним з напрямів досліджень семантики є ідеографічний опис лексики та класифікації мовних даних. До нього звичайно застосовують два основних підходи: індуктивний, який полягає у моделюванні семантичних відношень у лексиці, представлених у вигляді ієрархії (від більш загального до часткового) або у вигляді семантичної мережі, в якій відсутнє мотивоване розташування лексем, і дедуктивний, який базується на апріорній класифікації понять, наприклад, у сфері філософії, мовознавства, математики тощо, а лексикографу залишається лише адаптувати її до мовного матеріалу. У наших дослідженнях із семантичної класифікації лексики ми застосували обидва принципи: індуктивний – для побудови ІПС термінів і дедуктивний – для побудови ідеографічного словника української публіцистики [2]. Методично ми розділили ці завдання, оскільки інформаційно-пошукові тезауруси (ІПТ) є моделлю семантики галузі знання, а ідеографічний словник містить інформацію про загальномовну лексику. Одиницею опису ТЗ є слово, а ідеографічного словника – поняття, які відображають класи суспільно значимих сутностей, розрізнюваних людьми.

Для більшості ідеографічних словників, тезаурусів, які активно почали укладатися з середини XX ст., характерним було прагнення їх розробників знайти і теоретично обгрунтувати якусь одну систематизацію лексики. Але практика показала, що через складність мовних об'єктів і зв'язок з мовною картиною світу мовців створення універсальної систематизації лексики у вигляді ієрархічної мережі відношень, яка може бути «накладена» на будь-яку мову, не можливо. Однак пошуки у цьому напрямку не припиняються.

Одним з актуальних міжгалузевих завдань нашого часу є логіко-понятійне моделювання терміносистем, необхідне при укладанні термінологічних словників, інформаційних тезаурусів, створенні автоматизованих інформаційних систем, баз даних, систем штучного інтелекту [1, с.63]. Частковим випадком моделювання знань можна вважати побудову інформаційно-пошукового тезауруса (ІПТ), який з одного боку, є способом формалізованого представлення термінології, тому що досить строго представляє семантичні відношення між термінами, а з іншого – вважається важливим джерелом постійного вдосконалення систем знань конкретних наук. Актуальність тезаурусотворення полягає в необхідності створення термінологічних банків даних (ТБД) для автоматичного семантичного аналізу науково-технічних текстів і термінологічних банків знань (ТБЗ) для накопичення значного обсягу термінологічної інформації.

В Україні проблемі тезаурусотворення не приділяється належної уваги, можна констатувати відсутність інформаційних технологій, які передбачають створення тезауруса як способу систематизації термінології і як інструменту інформаційного пошуку. Саме ця обставина була зовнішнім стимулом для цього дослідження. Внутрішнім стимулом був позитивний досвід, набутий при створенні інформаційно-

пошукової системи лінгвістичної термінології – теми, підтриманої Департаментом з інноваційних технологій при Міністерстві освіти і науки України.

Мета дослідження: 1) укладання електронного Словника літературознавчих термінів з використанням формалізованої методики конструювання тезауруса, що відповідає сучасним стандартам термінографії, та представлення його в мережі Інтернет; 2) верифікація теоретичної тезаурусної моделі шляхом застосування її для аналізу корпусу літературознавчих текстів, викладених в Корпусі текстів української мови [www.MOVA.info.].

На першому етапі створювалася інформаційно-пошукова система (ІПС) у вигляді електронної бази літературознавчих термінів (в її створенні брали участь студенти-магістри спеціалізації «комп'ютерна лінгвістика» Інституту філології; програмне забезпечення – В. Сорокін). В алфавітному словнику для кожного слова-терміна (більше 1350 од.) надається тлумачення. Тезаурусний словник, крім алфавітного списку термінів і тлумачної частини до кожного терміна, містить перелік логіко-семантичних відношень між літературознавчими термінами (список запозичено з роботи [5], але доповнено і модифіковано нами). Розроблена ІПС включає не тільки множину окремих термінів, представлених у вигляді алфавітного списку з їхніми тлумаченнями, а й самі моделі представлення зв'язків між термінами.

На основі літературознавчих словників [3; 4] в компактній та доступній формі подано тлумачення термінологічних одиниць. До словника включено літературознавчі терміни, переважно іменники або іменникові словосполучення. Відбір словникових одиниць до бази даних здійснювався на евристичних засадах (знання укладачів тезауруса, експертів-літературознавців)) (Рис. 1).

Рисунок 1. Фрагмент електронного словника літературознавчої термінології

Для усвідомлення понятійних зв'язків між літературознавчими термінами кожному терміну надається тлумачення (без посилання на джерела, хоча джерелами слугували авторитетні літературознавчі термінологічні словники, енциклопедія, а у перспективі монографії, підручники, щоб охопити якомога більшу кількість термінологічної лексики, зокрема новотворів). Це пояснюється тим, що першим способом компресії наукового знання, одним з видів семантичного представлення терміна є його дефініція. Відомо, що значення терміна у вигляді дефініції ніколи не передає всього змісту поняття, але воно відображає важливі ознаки наукового поняття, тому є джерелом формування зв'язків аналізованого терміна з іншими термінами досліджуваної терміносистеми.

Побудова тезауруса (ТЗ) передбачає розкриття всіх типів відношень між термінами, основними з яких є гіпонімія (рід-вид), супідрядність на одному рівні – парціація (частина-ціле), синонімія, кореляція, асоціація, локалізація об'єкта, його призначення, функція, способи вираження функції тощо. Зміст відношень розширено настільки, щоб можна було охопити максимально широкий пласт термінів, з якими зв'язаний аналізований термін як реєстровий. Оскільки зміст тлумачення був недостатнім для здобуття всіх істотних для термінів відношень, ми орієнтувалися на енциклопедичні словники, наукові праці з проблематики літературознавства, знання свої власні та літературознавців-фахівців.

Словникова стаття побудована у вигляді анкети, «пропонованої» кожному терміну. В анкеті вміщено стандартний перелік відношень, які щодо реєстрового слова є поняттєвими. Назва відношення є двомісним предикатом R(A,B), який зв'язує заголовне слово тлумачної статті (А) і введений цим предикатом термін (В) [5, с.22].

Тезаурус містить 1350 термінів, які охоплені семантичною мережею із 4163 семантичних відношень (табл. 1).

Табл. 1. Кількісні характеристики семантичних відношень

Тип семантичних відношень (R)	Термін А (приклад)	Термін В (приклад)	Частота реалізації семантичних відношень у ТЗ
1	2	3	4
Дивись…(про А дивись В)	ударник	Акцентний вірш	596
Асоціація (А асоціюється з В)	металогія	алегорія	544
Рід – Вид (А є родовим до В)	акромонограма	римована акромонограма	494
Об'єкт науки (А є об'єктом В)	Астрофічність вірша	поетика	432

Продовження табл. 1

1	2	3	4
Синоніми (А синонімічний В)	акромонограма	лексико-композиційний прийом	312
Аспект (А розглядається в аспекті В)	пастораль	антична поезія	299
Параметр (А характеризується В)	жонглерська поезія	алітерація	259
Спосіб представлення об'єкта (А представляється через В)	асонанс	евфонія	170
Частина – Ціле (В складається з А)	рядок	баяті	161
Корелят (А протилежний В)	каталектика.	Акаталектика	150
Дисципліна (А розглядається в дисципліні В)	атонування	ритміка	143
Локалізація (А локалізується в В)	бейт	арабомовна писемна поезія	113
Носій параметра (носієм параметра А є В)	ашуг	баяті	91
Інструмент/Метод (А із застосуванням В)	бестіарій	алегорія	84
Спосіб вираження (А виражається В)	александрійський вірш	шестистопний ямб	56
Жанр (А є жанром В)	пастораль	антична поезія	50
Функція основна (А виражає В)	атонування	втрата акценту	36
Рівень мови (А розглядається на рівні, позначеному В)	еліпс	синтаксичний рівень	33
Об'єкт кінцевий (А здійснюється над В)	атонування	слово	28
Клас (А входить до класу В)	мадригал	альбомна поезія	24
Розмір (А характеризується (вимірюється) В)	білий вірш	Метричний стопний вірш	24

Закінчення табл. 1

1	2	3	4
Операція/процедура (для А процедурою / операцією буде В)	астрофічність вірша	Деструкція	18
Об'єкт початковий (В здійснюється над А)	метричний стопний вірш	атонування	17
Імплікація (якщо А, то В)	бурлеск	комічна поезія	16
Лінгвістичний об'єкт (А представляється у вигляді В)	віршований твір	буріме	9
Відноситься до…(стосується термінів-ад'єктивів: А відноситься до В)	Безсполучниковий	безсполучниковість	4

Всього встановлено 26 семантичних відношень. Як видно з таблиці, для цієї терміносистеми найчастотнішими виявилися непрямі парадигматичні відношення «дивись…» (596 реалізацій), що вказує на терміни, понятійно близькі між собою, й «асоціація» (544), які охоплюють разом 27,4% зв'язків між термінами (від 4163 – загальної кількості семантичних відношень у терміносистемі). Найголовніші прямі парадигматичні семантичні відношення (рід-вид, синонімія, частина-ціле, кореляція) охоплюють майже таку ж частку всіх парадигматичних відношень (26%). З точки зору теоретичної семантики, чим більше у словнику міститься семантичної інформації, тим краще.

Оскільки словникова стаття представляє собою синтез інформації літературознавчої, тлумачної, енциклопедичної, що представлено формалізовано у зв'язку з інформаційним підходом, розрахованим на запит користувача, можна створити діалогову систему на зразок:

Запит: «З якими ще термінами і якими відношеннями пов'язаний термін буф?

Відповідь: «п'єса, драматичний твір» – відношення вид-рід; «трагедія» – відношення бути корелятом; «комічна п'єса, гротеск» – відношення дивись; «комічна п'єса» – відношення синонімії; «теорія літератури, літературознавство» – відношення дисципліна.

Процедурно відповіді на запит здобуваються з тезаурусного графа у вигляді семантичної мережі, яка представляє собою ієрархічно організовану структуру даних – термінів – вузлів і дуг, що виражають різні типи тезаурусних відношень і автоматично будуються з тезауруса у текстовому вигляді.

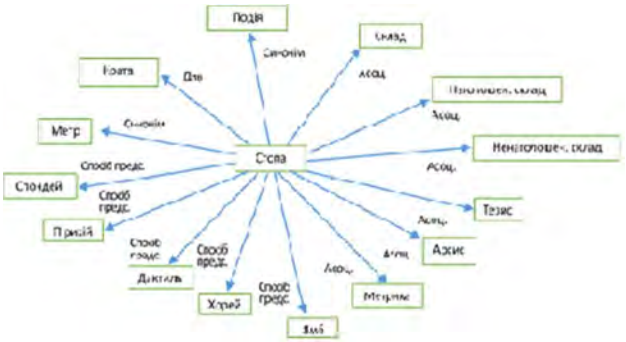

Рис 2. Граф тезаурусних зв'язків терміна «стопа» (словесно-символьне графічне зображення)

На рисунку представлено фрагмент графа, де стрілки спрямовані тільки від «хазяїна» (реєстровий термін стопа) до «слуги». (В Інтернеті на порталі у рубриці «Тезауруси» унаочнено ще один інтерфейс тезаурусних графів словесного типу, де чорним кольором на екрані комп'ютера позначаються відношення «хазяїн (х.) – слуга (с.)», коли реєстровий термін є «хазяїном», підпорядковувальним, а синім – «слуга – хазяїн», тобто, навпаки, є слугою, отже, підпорядкованим.

Оскільки кожне з відношень має свою внутрішню службову мітку і номер в електронному форматі, це дозволяє одержувати дані, згруповані так, як це потрібно користувачеві. Представлену модель ТЗ можна розглядати як семантичний та інформаційний опис лінгвістичної термінології, яка модифікується в: а) алфавітний словник із дефініціями; б) словник синонімів; в) словник родо-видових відношень; г) автоматичний інформаційний довідник тощо.

Наступний етап цього проекту – побудова динамічної логіко-понятійної моделі кожного конкретного тексту з корпусу текстів літературознавчої тематики.

Методика укладання тезауруса на конкретному тексті полягає у:

1) лематизації та впорядкуванні слів тексту за частиномовною приналежністю;

2) визначенні для кожної леми (іменникової або прикметникової) абсолютної частоти вживання;

3) побудові тезаурусного графа термінів конкретного тексту з абсолютними частотами вживання у тексті шляхом накладання тезаурусного графа терміносистеми;

4) знятті омонімії значень термінів через звертання до контексту;

5) побудові допоміжного реєстру слів з абсолютними частотами, які не увійшли до тезаурусу;

6) пошуку слів-претендентів на терміни і фіксації їх з ілюстративним контекстом у додатковому термінологічному реєстрі.

Пропонована методика дає можливість в автоматизованому режимі виявити, а потім вивчити термін у сфері функціонування.

Тезаурус ілюструє семантичну безперервність: у словнику немає і не може бути термінів, ізольованих в семантичному відношенні. Від семантичної мережі можна перейти й до тематичної. Темами є назви тематичних груп тезаурусного дерева, з іншого боку, тематична мережа становить ієрархію провідних і додаткових тем. Вихідними будуть такі методичні положення: 1) тематично значимі термінологічні групи будуються на основі тезаурусного дерева; 2) ієрархія тем тексту визначається залежно від кількості термінів, які заповнили тематичні групи.

Розроблений текстозорієнтований ІПТ із термінологічним апаратом і програмним забезпеченням може бути використаний при створенні термінологічних баз даних (ТБД) і термінологічних баз

знань (ТБЗ). Переваги такого словника перед паперовим у тому, що електронний ІПТ допускає вхід через будь-яку характеристику терміна. З іншого боку, термін не обмежений рамками своєї «словникової статті», як у паперовому словнику – в ТБД встановлюються логічні й асоціативні зв'язки між термінами.

Важливість цього дослідження полягає в тому, що, по-перше, ІПС у мультимедійному просторі забезпечить широке коло літературознавців сучасним стандартизованим словником літературознавчих термінів; по-друге, здобутком Проекту є методика конструювання тезаурусів, а також комп'ютерний інструментарій для реалізації цієї методики; по-третє, ІПС сумісний з інтелектуальними системами опрацювання текстової інформації, в яких він може бути використаний як база знань та інструмент розпізнавання смислу текстів (автоматичне реферування й анотування наукового тексту літературознавчої тематики), по-четверте, створено класифікаційну модель наукового знання – вона включає термінологічний словник і лінгвістичне забезпечення конкретної ІПС. Якщо будуть створені за такою методикою ІПС різних, наприклад гуманітарних, наук, можна створювати автоматично семантичну мережу – модель плану змісту, яка є моделлю семантики галузі гуманіратного знання.

Література

1. Герд А.С. Прикладная лингвистика / А. С. Герд. — СПб. : Изд-во С.-Петерб. ун-та. — 2005. — 266, [1] с.
2. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи: монографія / Н. Дарчук. — К. Освіта України, 2013 — 543 с.
3. Квятковский А.П. Поэтический словарь / Квятковский А. П. — М.: Сов. Энцикл., 1966.
4. Літературознавчий словник-довідник / Р.Т. Гром'як, Ю.І. Ковалів та ін. – К.: ВЦ «Академія», 1997. – 752 с. — Режим доступу: http://shron.chtyvo.org.ua/Hromiak_Roman/Literaturoznavchyi_slovnyk-dovidnyk.pdf
5. Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике / С. Е. Никитина. — М. : Наука, 1978. — 375 с.