УДК 81’37

Галина Мацюк,
викладач Тернопільського національного технічного університету ім. Івана Пулюя (м. Тернопіль, Україна)

вул. Руська, 56, м. Тернопіль, 46001, Україна
E-mail: galuna.matsiuk@gmail.com

Наталія Кунанець,
д-р наук із соціальних комунікацій, професор Національного університету «Львівська політехніка» (м. Львів, Україна)

вул. С. Бандери, 12, м. Львів, 79013, Україна
E-mail: nek.lviv@gmail.com

Представлення знань у вузькоспеціальних предметних областях за допомогою тезауруса

У статті рокривається роль галузевого інформаційно-пошукового тезауруса у процесах систематизації термінології предметної області. Інформаційно-пошуковий тезаурус розглядається як мовна модель певної предметної області. Аналізуються можливості застосування інформаційно-пошукового тезауруса як засобу підвищення ефективності пошуку в сучасних інформаційно-пошукових системах. Пропонується алгоритм формування тезаурусів для вузькоспеціальних предметних областей, обґрунтовується необхідність створення тезаурусної моделі предметної області.

К л ю ч о в і с л о в а: тезаурус, предметна область, інформаційний пошук, термін, дескриптор.

Стрімке зростання обсягів інформації – це одна з рис формування сучасного суспільства. Призупинити цей процес уже неможливо. Постійне розширення фронту наукових досліджень генерує збільшення масивів опублікованих наукових матеріалів. Проблемність орієнтації та пошуку у величезному масиві недостатньо упорядкованих матеріалів часто призводить до втрати часу, матеріальних та інтелектуальних ресурсів. Тому виникає нагальна потреба у вдосконаленні як систем опрацювання, зберігання інформації, так і навігації у цих системах. Один із шляхів вирішення окресленої проблеми – це систематизація неструктурованих даних. Подання знань у формалізованому вигляді не тільки значно прискорює процеси роботи з інформацією, а й підвищує продуктивність її пошуку.

Основним засобом підвищення повноти і точності пошуку є ефективно розроблене лінгвістичне забезпечення. Лінгвістичні засоби виступають інтерфейсом між природною мовою та формальними пошуковими алгоритмами інформаційно-пошукових систем. Лінгвістичне забезпечення формується з низки елементів. У першу чергу, це штучна мова подання даних у інформаційно-пошукових системах, яка визначає архітектуру, синтаксис, семантику подання інформації в базах даних інформаційно-пошукових систем та інформаційно-пошукова мова, тобто мова, якою користувач звертається до інформаційної системи для отримання відомостей, які його інтересують [4]. Саме лінгвістичне забезпечення гарантує ефективну реалізацію таких процесів, як індексування документів і запитів. Воно сприяє ефективному формуванню тематичних пошукових запитів у базах даних, налагоджує міжсистемну інформаційну взаємодію. Якщо ж між масивом документів і користувачем розмістити «посередника», тобто тезаурус, то це значно звузить обсяг отриманих результатів пошуку і підвищить рівень їх релевантності.

Розвиток інформаційних технологій генерує можливість інформаційного моделювання вузькоспеціальної предметної області і подання отриманих даних у формі тезауруса, який є гіпертекстом, що відображає ієрархічно організовану семантичну структуру певної предметної області.

Поняття «тезаурус» сьогодні активно використовується в галузях штучного інтелекту, в інформаційних технологіях та бібліотекознавстві. І це цілком зрозуміло, адже нині постає дедалі більше інтелектуальних завдань, пов’язаних з опрацюванням інформації: індексуванням документів, інформаційним пошуком і автоматичним аналітико-синтетичним опрацюванням документів, які вирішуються саме із застосуванням тезаурусів.

Мета статті – розкрити особливості формування тезауруса як засобу систематизації термінології

вузькоспеціальних предметних областей та засобу підвищення ефективності пошуку в сучасних інформаційно-пошукових системах.

Інноваційна галузь дослідження, яка отримала назву «Розумне місто», набирає стрімких темпів розвитку в усіх сферах суспільної діяльності. Такий стан речей призводить до розширення контактів, до глобалізації професійної комунікації, генерує нагальну потребу в створенні комплексного двомовного тезаурусу, який дає змогу окреслити єдині спільні межі термінологічного поля даної області на рівні декількох мов та полегшити дослідникам інформаційний пошук у даній предметній області.

Такий підхід є важливим кроком на шляху впорядкування, стандартизації лексики зазначеної області знань. Сформований тезаурус як результат лексикографічного опису даної предметної області виступає своєрідним нормативним документом кодифікованих термінів, котрі орієнтують користувачів на правильне їх застосування. Це, у свою чергу, сприяє поглибленню контактів і поліпшенню взаєморозуміння між фахівцями різних країн.

Історія виникнення тезаурусів бере свій початок у II–III ст. н. е. Саме тоді з'явився санскритський словник «Амарокоша», який містив близько 10 тис. слів і складався з трьох книг, кожна з яких ділилася на глави, глави – на секції.

Термін «тезаурус» у XIII ст. вперше використав флорентійський учений Брунетто Латіні у заголовку своєї праці – систематизованої енциклопедії «Книга про скарб», що цілком відповідало семантиці слова «thesauros», тобто «скарб», «багатство», «запас».

Сучасний етап історії ідеографічних словників відкривається роботою П. М. Роже «Тезаурус англійських слів і виразів» (Peter Mark Roget. «Thesaurus of English Words and Phrases». 1852 р.) [7]. Вчений спробував охопити весь лексичний фонд англійської мови. Всі елементи лексичної системи англійської мови він розбив на такі понятійні кластери: абстрактні відносини, простір, фізика, матерія, почуття, інтелект, воля, психологічні стани. Далі ці групи поділяються на 24 класи, класи – на підкласи, підкласи розпадаються на категорії, категорії – на секції, секції – на групи. Всього П. Роже виокремив більше тисячі понятійних груп, і кожна містила слова, близькі за змістом.

Варто зазначити, що перші тезауруси не враховували особливостей інформаційної діяльності, вони були пов'язані з фундаментальними проблемами пізнання, відображенням засобами природної мови сутності навколишнього світу та закономірностей його сприйняття.

Термін «тезаурус», пов'язаний з обчислювальними машинами, вперше вжила Анетта Мастерман. У 1954 р. вона звернула увагу на те, що тезаурус – досить зручний засіб для опису семантичних структур природної мови. Вона запропонувала використати цей засіб у технологіях машинного перекладу. У 1970-х рр. розпочався «тезаурусний бум», пов'язаний з розробленням інформаційно-пошукових систем. Відтоді термін «тезаурус» міцно узвичаївся як у лінгвістів, так і у фахівців з інформаційних технологій [1].

Проаналізувавши роботи В. В. Морковкіна [3], А. Я. Гладун [2], А. Н. Городішева [6], можна вивести таке визначення тезауруса. Тезаурус предметної області – система знань, подана у вигляді набору ключових термінів або дескрипторів цієї області, пов'язаних між собою певними семантичними відношеннями, що відображають основні співвідношення понять предметної області знань, котра описується. Основним призначенням тезауруса є підвищення ефективності пошуку необхідної інформації. Отже, тезаурус – це спосіб систематизації знань певної предметної області, інструмент ефективного інформаційного пошуку.

Тезауруси використовуються і як інструменти термінологічного контролю в процесі аналізу, індексування документів й інформаційних запитів, а також під час автоматизованого пошуку інформації. Функціональна роль тезауруса в інформаційно-пошукових системах висуває високі вимоги до якості підготовки тезауруса, від ступеня досконалості якого, як правило, залежить ефективність пошуку.

Тезаурус є різновидом загального або спеціального словника, який вирізняється тим, що методика розроблення та побудови його словникових статей дає змогу відобразити синонімічні, антонімічні, паронімічні, гіпогіперонімічні та інші семантичні відношення між лексемами. Таке комплексне відображення термінологічної системи предметної області робить тезаурус досить ефективним і еквівалентним помічником при описі певної предметної області. Саме тому словники тезаурусного типу набувають дедалі більшого поширення.

Характерна особливість тезауруса: статті в ньому побудовані таким чином, що відображають семантичні відтінки близьких за значенням слів. Якщо традиційний тлумачний словник дає тільки

визначення слова, то статті тезауруса розкривають значення слова як через дефініцію, так і через порівняння з іншими словами і понятійними групами, через експлікацію семантичних зв'язків одного терміна з іншими.

Головна відмінність тезауруса від традиційного словника полягає в тому, що він подає не лише компіляцію статей, а й систему термінів-дескрипторів. Ключові слова – еквівалентні і близькі за значенням, і саме за їх допомогою здійснюється опрацювання та пошук інформації. Вони об'єднуються в клас умовної еквівалентності. Кожен такий клас є словниковою одиницею, яка подається у формі окремого слова, словосполучення або коду. Така словникова одиниця і є дескриптором. Класифіковані за змістом дескриптори утворюють тезаурус.

Процес побудови тезауруса передбачає здійснення таких кроків:

1) попередній відбір лексичних одиниць (складання списків ключових слів, словників);
2) побудова класів умовної еквівалентності, тобто перетворення лексичних одиниць у задану стандартну форму;
3) встановлення наявних семантичних відношень.

Розроблення методики, відбір лексичного матеріалу, сам процес укладання тезауруса передбачають значний обсяг науково-дослідної роботи, пов'язаної з аналізом терміносистеми предметної області. Вона проводиться для того, щоб виявити терміни-претенденти до включення у тезаурус та їх окремі значення. При визначенні основних принципів упорядкування та стандартизації галузевої термінології визначальною є пара «поняття – термін», а на більш високому рівні: «система понять – система термінів».

Побудова тезауруса – це складний багатоопераційний процес. Кожен етап роботи пов'язаний з аналізом багатьох варіантів. Для того, щоб отримати новий інтелектуальний продукт, необхідно створити певні умови: чітке розуміння мети та постановки завдання; достатність інформаційної бази; повний опис об'єктів предметної області; чітку співпрацю групи дослідників.

Науково-дослідна робота з укладання тезауруса розпочинається з відбору лексичного матеріалу. Певні труднощі на цьому етапі пов'язані насамперед з необхідністю розмежувати термінологічну і загальновживану лексики. З накопиченням термінологічного матеріалу, що видобувається з текстів, визначаються основні термінологічні поля області, що досліджується.

Сучасні науковці докладають багато зусиль, щоб віднайти об'єктивні критерії, які допоможуть однозначно виокремити термінологічні одиниці, що виражають поняття і зв'язки з тією чи іншою галуззю знань, і у такий спосіб визначити обсяг і рамки термінології. Дослідники по-різному вирішують проблему виявлення термінів у тій чи іншій підсистемі, використовуючи при цьому різні визначення терміна, пропонуючи свої методики відбору термінів, критерії їх відбору у випадку, коли система понять ще не склалася і немає чітких граней між «терміном» і «не терміном» [5].

При побудові інформаційно-пошукового тезауруса першочерговим завданням постає відбір термінів для включення до нього. Існує декілька можливих джерел відбору термінів при розробленні тезауруса. Насамперед повинні бути вивчені тезауруси близьких предметних областей, котрі можуть містити значну кількість термінів, що мають увійти до структури нового тезауруса. Терміни-кандидати на внесення до тезауруса можуть пропонуватися експертами предметної області. Ще одним джерелом отримання термінологічних одиниць є наукові тексти.

Терміни, що включаються до тезауруса, мають відповідати таким вимогам:

- терміни тезауруса повинні передавати поняття, які присутні в наукових текстах, і вони мають відбиратися з міркувань ефективності їх використання в пошуку документів;
- важливим чинником включення терміна є частотність його вживання в текстах;
- включення нових термінів до тезауруса має відбуватися з урахуванням вже наявних термінів: потрібно перевіряти, чи представляє термін-кандидат окреме поняття, якому немає відповідників серед вже існуючих термінів тезауруса; необхідно уникати включення термінів, значення яких перетинаються зі значеннями вже існуючих термінів у тезаурусі.

Відбір термінів здійснюється на основі таких критеріїв:

- інформативність: можливість терміна-кандидата означувати мовне явище та одночасно бути основою для інформаційного запиту;
- частотність: відбираються терміни, які найбільше використовуються в предметній області;
- відповідність темі: відбираються терміни, які відображають проблему;
- актуальність: перевага віддається термінам, що відображають сучасні розробки в даній галузі наук або мають велике значення для її розуміння;

- практична значущість: виокремлюються терміни, розуміння яких справляє істотний вплив на з'ясування текстів.

Головна функція тезауруса – сприяти пошуку та відбору необхідної інформації. Проте структура й принципи організації, що забезпечують достатнє наповнення термінами і глибину розкриття семантичних взаємозв'язків, дають змогу тезаурусу моделювати термінологічну систему предметної області найбільш повно і системно. Практична цінність тезауруса предметної області зумовлюється можливістю використовувати його одночасно для аналізу і конструювання лексичної системи, класифікації і зберігання термінологічних даних, опрацювання інформації в пошукових системах.

Тезаурус репрезентує систему термінів відповідної галузі. За його допомогою можна не лише здійснювати пошук, а й вивчати певні поняття, отримувати систему відношень даного терміна з іншими термінологічними одиницями, з'ясувати його роль у системі знань певної галузі. Тезаурус впливає на мовну компетенцію фахівців, сприяє оновленню термінології певної області знань, робить доступними наукові джерела інформації, допомагаючи різномовним носіям зрозуміти один одного та сприяти поглибленому вивченню тих мов, якими представлений тезаурус.

Оскільки тезаурус є ефективним засобом представлення та систематизації знань певної предметної області, розроблення тезауруса предметної області дає змогу повніше представити її термінологічну систему з урахуванням різноманітності семантичних зв'язків.

Подальші дослідження у цьому напрямі мають уніфікувати терміносистему предметної області «Розумне місто», ґрунтуючись на аналізі корпусу різних джерел, у яких розкриваються терміни галузі, що досліджується. Виділення термінів з контекстів і проведення семантичного аналізу термінів, а також дефініціації ключових термінів уможливить представлення всієї термінології області тезаурусним методом.

Тезаурус забезпечує систематизоване уявлення термінології в області «Розумного міста». Він дає змогу сформулювати термінологічну систему, створити базу для систематизації знань та інформаційних ресурсів і забезпечити зручний доступ до них. Двомовність тезауруса дозволяє вітчизняним науковцям і фахівцям швидше та ефективніше орієнтуватися у предметній області.

Список бібліографічних посилань

1. Андреев А. М., Березкин Д. В., Симаков К. В. Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. URL: http://rcdl.ru/doc/2004/paper14.pdf

2. Гладун А. Я., Рогушина Ю. В. Основи методології формування тезаурусів з використанням онтологічного та мереологічного аналізу. *Штучний інтелект.* 2008. № 4. С. 53–61.

3. Морковкин В. В. *Идеографические словари.* М.: Изд-во МГУ, 1970. 70 с.

4. Олифер В., Олифер Н. *Основы компьютерных сетей.* СПб, 2009.

5. Шуневич Б. І. Сучасні способи відбору термінів та укладання перекладних словників нових терміносистем. *Вісник Житомирського державного університету імені Івана Франка.* 2008. Вип. 38. С. 90–93.

6. Gorodishcheva A. N., Gorodishchev A. V. Standards, specifications and features of the logical linguistic terminology management systems, communications. *Science Prospects.* 2011. № 5 (20). P. 62–65.

7. Peter Mark Roget Thesaurus of English Words and Phrases (Fourth ed.). London: Longman, Brown, Green, and Longmans. 1856. URL: https://books.google.com.ua/books?id=9nYCAAAAQAAJ&printsec=frontcover&hl=uk#v=onepage&q&f=false (Last accessed: 15.07.2018).

References

1. Andreev, A. M., Berezkin, D. V., Simakov, K. V. (2018). Osobennosti proektyrovaniia modeli i ontolohii predmetnoi oblasty dlia poiska protyvorechii v pravovykh elektronnykh bibliotekakh [Features of the design of the model and ontology of the subject area for the search for contradictions in legal electronic libraries]. Retrieved from http://rcdl.ru/doc/2004/paper14.pdf [In Russian].

2. Hladun, A. Ya., Rohushyna, Yu. V. (2008). Osnovy metodolohii formuvannia tezaurusiv z vykorystanniam ontolohichnoho ta mereolohichnoho analizu [Fundamentals of the thesaurus creation methodology using ontological and mereological analysis]. *Shtuchnyi Intelekt, 5,*112-124. [In Ukrainian].

3. Morkovkin, V. V. (1970). *Ideograficheskie slovari* [Ideographic dictionaries], Moscow, Russia: Izd-vo MHU. [In Russian].

4. Olifer, V., Olifer, N. (2009). *Osnovy kompiuternykh setei* [Basics of computer networks]. St. Petersburg, Russia. [In Russian].

5. Shunevich, B. I. (2008). Suchasni sposoby vidboru terminiv ta ukladannia perekladnykh slovnykiv novykh terminosystem [Modern ways of selecting terms and making translated dictionaries of new terminology systems]. *Visnyk Zhytomyrskoho Derzhavnoho Universytetu Imeni Ivana Franka, 38,* 90-93. [In Ukrainian].

6. Gorodishcheva, A. N., Gorodishchev, A. V. (2011). Standards, specifications and features of the logical linguistic terminology management systems, communications. *Science Prospects, 5* (20), 62-65.

7. Peter Mark Roget. (1856). Thesaurus of English Words and Phrases (Fourth ed.). London: Longman, Brown, Green and Longmans. Retrieved from https://books.google.com.ua/books?id=9nYCAAAAQAAJ&printsec=frontcover&hl=uk#v=onepage&q&f=false

Halyna Matsiuk,
Lekture of Ternopil Ivan Puliui National Technical University (Ternopil, Ukraine)

Nataliia Kunanets,
Doctor in social communications, Professor of the Department of Information Systems and Networks,
Lviv Polytechnic National University (Lviv, Ukraine)

REPRESENTATION OF KNOWLEDGE IN NARROW FOCUSED DATA DOMAINS USING THESAURUS

The role of the branch information retrieval thesaurus in the processes of data domain terminology systematization is analyzed in this paper. Information retrieval thesaurus is considered as the language model of a certain data domain. The possibility of using the information retrieval thesaurus as the means of increasing the retrieval ratio in modern information retrieval systems is considered. The algorithm of thesaurus formation for narrow focused data domains is offered. The importance of creating the data domain thesaurus model is substantiated in this paper.

K e y w o r d s: thesaurus, subject area, information search, term, descriptor.

Ternopil Ivan Puliui National Technical University,
56, Ruska str., Ternopil, 46001, Ukraine
E-mail: galuna.matsiuk@gmail.com
12, ul. S. Bandery, Lviv, 79013, Ukraine
E-mail: nek.lviv@gmail.com

Стаття надійшла до редакції 21.01.2019 р.