УДК 002

DOI: https://doi.org/10.32461/2409-9805.2.2019.175881

Мацюк Галина Ростиславівна,
викладач Тернопільського національного
технічного університету ім. Івана Пулюя
galuna.matsiuk@gmail.com
http://orcid.org/0000-0002-8857-1857

ІНФОРМАЦІЙНО-ПОШУКОВІ ТЕЗАУРУСИ: СВІТОВИЙ ТА ВІТЧИЗНЯНИЙ ДОСВІД ФОРМУВАННЯ

Мета роботи. *Дослідити основні види, змістовне наповнення та кількісні показники інформаційно-пошукових тезаурусів у світовій та вітчизняній практиці.* ***Методологія дослідження.*** *У процесі дослідження застосовано наукові методи аналізу, синтезу, узагальнення.* ***Наукова новизна*** *роботи полягає у проведені аналізу функціонуючих інформаційно-пошукових тезаурусів у світовій та вітчизняній практиці.* ***Висновки.*** *Проведений аналіз існуючих інформаційно-пошукових тезаурусів, принципів їх побудови і функціонування у різних сферах, дає підстави зробити висновок, що специфіка предметної області кожного тезауруса знаходить відображення в його структурі.*

Ключові слова: *тезаурус, термін, дескриптор, аскриптор.*

Мацюк Галина Ростиславовна
преподаватель Тернопольского национального
технического университета им. Ивана Пулюя

ИНФОРМАЦИОННО-ПОИСКОВЫЕ ТЕЗАУРУСЫ: МИРОВОЙ И ОТЕЧЕСТВЕННЫЙ ОПЫТ ФОРМИРОВАНИЯ

Цель работы. *Исследовать основные виды, содержательное наполнение и количественные показатели информационно-поисковых тезаурусов в мировой и отечественной практике.* ***Методология****. В процессе исследования применены научные методы анализа, синтеза, обобщения.* ***Научная новизна*** *работы заключается в проведении анализа функционирующих информационно-поисковых тезаурусов в мировой и отечественной практике.* ***Выводы****. Проведенный анализ существующих информационно-поисковых тезаурусов, принципов их построения и функционирования в различных сферах, дает основания сделать вывод, что специфика предметной области каждого тезауруса находит отражение в его структуре.*

Ключевые слова*: тезаурус, термин, дескриптор, аскриптор.*

Matsiuk Halyna,
Lekture of Ternopil Ivan Puluj National Technical University

INFORMATION RETRIEVAL THESAURI: WORLD AND NATIONAL EXPERIENCE

Purpose of Article*. The objective of the paper is to explore the main types, content and quantitative indicators of information retrieval thesauri in the world and. national experience.* ***Methodology****. Methods of analysis, synthesis, and generalization are applied in the research process.* ***The scientific novelty*** *of this paper is the analysis of functioning information retrieval thesauri in the world and national experience.* ***Conclusions.*** *The analysis of existing information retrieval thesauri, the principles of their construction and functioning in various spheres provides the reasons to conclude that the specificity of the date domain of each thesaurus is reflected in its structure.*

Key words*: thesaurus, term, descriptor, non-descriptor.*

Актуальність теми дослідження. В умовах розвитку сучасного інформаційного суспільства користувачам для того, щоб орієнтуватися у накопичених великих фондах і потужному потоці інформації, знаходити дані, що відповідають їхнім запитам, необхідні допоміжні засоби, якими є спеціальні методи опрацювання даних, організації швидкого та ефективного пошуку. Для вирішення цієї проблеми призначені інформаційно-пошукові тезауруси.

Мета дослідження полягає в аналізі основних видів та змістовного наповнення інформаційно-пошукових тезаурусів, які використовуються у світовій та вітчизняній практиці, як інформаційного ресурсу репрезентації термінологічного апарату та структури певної предметної області та інструменту формування пошукових образів та пошукових запитів.

Виклад основного матеріалу. Розробленя інформаційних ресурсів для ефективного формування пошукових образів та пошукових запитів через постійне зростання інформаційних набуває все більшої актуальності. Одним з найвідоміших інформаційних ресурсів для реалізації цієї мети є універсальний комп'ютерний тезаурус WordNet [30], створений командою Принстонського університету США під керівництвом психолінгвіста Джорджа Міллера. Це велика база даних англомовних лексичних одиниць. Онлайн-версія WordNet 3.1. містить 155 327 слів, організованих у 175 979 синсетів, та 207 016 пар лексем. Особливість цього тезаурусу в тому, що базовою одиницею є не окреме слово, а синонімічний ряд – синсет (synset), який об'єднує слова зі схожим значенням, які і є вузлами мережі. Для зручного використання тезаурусу кожен синсет доповнений дефініцією і прикладами вживання слів в контексті. Синсети взаємопов'язані за допомогою концептуально-семантичних і лексичних відношень: гіперонім – hyperonymy (breakfast → meal), гіпонім – hyponymy (meal → lunch), має ієрархію підпорядкування – has-member (faculty → professor), окремі одиниці – member-of (pilot → crew), мероніми – meronymy (table → leg) та ін. Тезаурус WordNet доступний он-лайн.

Концепція WordNet-у лягла в основу створення ряду тезаурусів аналогічної тематики іншими мовами. Серед них тезаурус EuroWordNet [14], який об'єднав терміни європейськими мовами, проте робота над його створенням була завершена влітку 1999 року. Дизайн бази даних, визначені відношення, топ-онтологія та інтерактивний індекс зараз заморожені. Разом з тим, завдяки створенню цього тезаурусу уперше була реалізована ідея об'єднання окремих термінологічних мереж у загальну систему.

Відповідно до засадничих принципів WordNet розроблений тезаурус термінів німецькою мовою GermaNet [10]. Версія GermaNet 13.0, розроблена станом на 2018 рік, містить 128100 синсетів, 164814 лексичних одиниць, 141774 концептуальних відношень. На відміну від WordNet, в якому застосований принцип роздільного опису різних частин мови та не прослідковувалися відношення між ними, у тезаурусі GermaNet представлено відношення між частинами мови.

За такими ж принципами було створено тезаурус DanNet [11], що містить 65 тис. синсетів датською мовою, MultiWordNet [21] для об'єднання термінів італійською мовою, а також низка національних тезаурусів, що містять концепти шведською, норвезькою, грецькою, португальською, баскською, каталонською, румунською, литовською, російською, болгарською, словенською мовами.

Тезауруси структуровані на тих самих засадах, що і американський WordNet для англомовних термінів (Princeton WordNet), проте кожен з них є унікальною мовною системою.

У 2001 р. була створена Всесвітня Асоціація WordNet (Global WordNet Association), завдання якої полягає в об'єднанні інформаційних ресурсів тезаурусного типу, розробленні загальних стандартів, що сприяють реалізації моделі WordNet для різних мов.

Використовуючи загальний принцип побудови WordNet було побудовано аналогічний російськомовний тезаурус RussNet [23], укладений дослідницькою групою Санкт-Петербурзького державного університету під керівництвом І. В. Азарової. Даний ресурс складається з чотирьох взаємозв'язаних частин, які сформовані з іменників, дієслів, прикметників та прислівників, які пов'язані між собою парадигматичними та синтагматични-

ми відношеннями. Тезаурус RussNet є найбільшим проектом подібного типу.

До того ж класу, що і WordNet належить російськомовний тезаурус РуТез [24], розроблений на базі словника суспільно-політичної тематики. Тезаурус РуТез є лінгвістичним ресурсом концептуального типу, тобто є ієрархічною мережею понять, до яких під'єднані текстові вирази. При цьому, на відміну від WordNet, який створювався як модель системи понять, присутніх у людській пам'яті (роздільний опис частин мови, спеціальні типи відношень та ін.), тезаурус РуТез розроблений саме як ресурс для автоматизованого опрацювання текстів. Даний ресурс містить термінологію економічної, політичної, військової, фінансової, законодавчої, соціальної, культурної та інших галузей діяльності. За принципами розроблення тезаурус РуТез об'єднує три способи розроблення комп'ютерних лінгвістичних ресурсів: традиційних інформаційно-пошукових тезаурусів; лінгвістичних ресурсів типу WordNet; формальних онтологій [2]. Нова версія тезауруса РуТез 2.0 містить 31.5 тис. понять, 111.5 тис. різних текстових входів (слів та виразів російської мови) і є інформаційним ресурсом закритого типу.

У Національному університеті «Львівська політехніка» було зроблено спробу розроблення тезаурусу, подібного до WordNet, що містить концепти українською мовою. У результаті дослідження було розроблено перші фрагменти української версії WordNet, в якому реалізовано 194 синсети, з яких 183 пов'язані гіпогіперонімічними, 14 антонімічними зв'язками, а також 150 зв'язками меронімії/голонімії та 6 also_see [19], що дає підстави стверджувати важливість продовження роботи над розробленням цього інформаційного ресурсу.

На сьогодні є понад 40 тезаурусів, розроблених різними мовами на основних засадах побудови WordNet, що демонструє ефективність інформаційних ресурсів такої структури.

Міжнародною Організацією Об'єднаних Націй з питань освіти, науки і культури, ЮНЕСКО (United Nations Educational, Scientific and Cultural Organization, UNESCO) створено багатомовний тезаурус ЮНЕСКО [31], укладений англійською, іспанською, французькою та російською мовами. Він містить 7 тис. термінів з галузей освіти, культури, природничих, соціальних та гуманітарних наук, комунікації та інформації, за ієрархічною класифікацією. Тезаурус складається із семи тематичних областей, розбитих на мікро масиви, що дозволяють швидко ознайомитися з темами: Освіта, Природничі науки, Культура, Соціальні та гуманітарні науки, Інформація та комунікація, Політика, право і економіка, Країни і групи країн. Кожен термін вищого рівня супроводжується посиланням UF (Used For) та спадковою ієрархією описів, перед кожним з яких знаходиться символ NT (Narrower Term – «вузький термін»). Для розкриття кожного терміна використовуються наступні позначення: SN (Scope Note – «зміст терміна»), MT (Microthesaurus – «мікротезаурус»), UF (Used For – «синонім»), BT (Broader Term – «широкий термін»), NT (Narrower Term – «вузький термін»), RT (Related Term – споріднений термін»). Тезаурус постійно оновлюється та поповнюється.

Видавничим бюро Європейського Союзу підтримується тезаурус EUROVOC [12]. Багатомовний інформаційно-пошуковий тезаурус EUROVOC відіграє роль міжнародного термінологічного стандарту. Він був спеціально розроблений для роботи з документною інформацією, яка зберігається в інституціях Європейського Союзу. EUROVOC з 1984 р. використовується для індексування та пошуку даних в інформаційно-пошукових системах Європейського парламенту, Бюро офіційних публікацій ЄС, парламентів європейських країн.

Бібліотеки інститутів Європейського Союзу, служби баз даних і їх користувачі використовують EUROVOC як довідкову термінологічну базу даних. Інституції Європейського Союзу використовують EUROVOC для створення пошукових образів для документації, що зберігається у створених ними базах даних: Європейський Парламент у EPOQUE, яка сформована з матеріалів парламентських документів і електронного каталогу з метаданими на них; Відділ офіційних публікацій ЄС – у каталогах системи CATEL, Офіційний журнал ЄС – у електронному архіві випусків.

Тезаурус EUROVOC містить понад 6000 еквівалентів дескрипторів, продубльованих 23 мовами та поєднаних зв'язками, характерними для кожної з мов. Інформаційно-пошуковий

тезаурус EUROVOC побудовано за ієрархічним принципом від загального до часткового. Верхній рівень складається із 21 розділу, яким присвоєно двозначний код та назву на природній мові, наприклад, 36 SCIENCE (НАУКА). Наступний рівень містить інформаційні масиви, розподілені на тематичні мікротезауруси (127). Кожен мікротезаурус ідентифікується чотиризначним кодом (перші дві цифри співпадають з кодом розділу, до складу якого входить мікротезаурус) та назвою на природній мові, наприклад, 3606 natural and applied sciences (3606 Природничі та прикладні науки). З огляду на можливі зміни в EUROVOC, розробники пронумерували розділи та мікротезауруси не суцільною нумерацією, а залишили лакуни в номерах: 04, 08, 10, …, 76. Числові коди розділів та мікротезаурусів однакові для усіх мовних версій.

Основні розділи тезауруса EUROVOC пов'язані з різними видами діяльності органів, установ та інститутів ЄС: політика, міжнародні відносини, право, економіка, торгівля, фінанси, соціальні питання, освіта та комунікація, наука, бізнес, транспорт, навколишнє середовище, промисловість, географія тощо.

У грудні 2018 року було опубліковано версію EUROVOC 4.9 [13], яка доступна для завантаження у форматі XML, SKOS-Core, SKOS-AP-EU (RDF), Marc-XML та Excel.

Науково-дослідним центром правової інформатики НАПрН України (під керівництвом члена-кореспондента НАПрН України Швеця М. Я.) у співпраці з Інститутом законодавства Верховної Ради України та Управлінням комп'ютеризованих систем ВР України у період з жовтня 2003 р. по червень 2009 р. проводилися дослідження, спрямовані на розроблення електронної інформаційно-пошуковий тезаурус EUROVOC 4.2, що містить дескриптори українською мовою. Українська версія тезаурусу також містить: 21 розділ; 127 мікротезаурусів, проте розширено до 6439 кількість дескрипторів (з них 511 верхнього рівня). Ця версія тезаурусу відображає 6448 ієрархічних та 3501 асоціативних зв'язків. На перших етапах робота над українською версією інформаційно-пошукового тезаурусу EUROVOC 4.2 спрямовувалася на досягнення відповідності української версії офіційним версіям країн ЄС, проте надалі для подолання обмежень щодо використання тезауруса при індексації термінів національної законодавчої бази планувалося використання аскрипторів та приміток і це передбачало формування версії тезаурусу EUROVOC 4.3 та 4.4. Проте, для реалізації цих задумів не було виділене фінансування, що призвело до припинення досліджень. Також, незважаючи на розроблене спеціалізоване програмне забезпечення, тезаурус EUROVOC 4.2 не використовується для аналітико-синтетичного опрацювання законів України та інших нормативно-правових актів, внесенні до бази даних «Законодавство України» їх повних текстів, доступ до яких є відкритим через веб-сайт zakon.rada.gov.ua [3].

Розроблення тезаурусів відбувається і в медичній галузі. Так, Національною службою охорони здоров'я Великобританії та Колегією американських патологів (College of American Pathologists) розроблено багатомовний тезаурус SNOMED CT [25], що містить систематизовану медичну номенклатуру. До складу SNOMED CT входять медичні терміни (terms), коди термінів (codes) і визначники кодів (definitions). У 2011 році до складу номенклатури SNOMED CT входить 311 000 концептів. Кожен з концептів має унікальний числовий ідентифікатор. Так, терміну «інфаркт міокарда» відповідає «22298006», «застуда» – «82272006» та ін.

Функціонує SNOMED CT для американського варіанту англійської мови, британського варіанту англійської мови, іспанської, датської, шведської, французької і нідерландської мов. Застосовується в медичній документації і звітах для підвищення ефективності роботи з клінічними даними.

Національною Бібліотекою Медицини Національного Інституту Здоров'я (США) (National Library of Medicine, National Institutes of Health (U.S.)) сформований тезаурус Medical Subject Headings або MeSH (Медичні предметні рубрики) [20], який використовується для індексування, каталогізації та пошуку інформації в галузі біології, медицини, охорони здоров'я та суміжних науках. Мовою оригіналу тезаурусу є англійська, проте існують національні версії MeSH іншими мовами, україномовної версії тезауруса MeSH не існує.

Станом на 2015 рік база даних MeSH містить 27455 головних рубрик, а також понад 220 тисяч додаткових термінів для полегшення пошуку. Тезаурус MeSH доступний он-лайн, для перегляду та безкоштовного завантаження.

SNOMED-CT та MeSH входять у масштабний уніфікований термінологічний та онтологічний ресурс, розроблений для інформаційних систем аналізу медичних текстів UMLS (Universal Medicine Language System).

Центральною науковою медичною бібліотекою Московської медичної академії ім. І.М. Сєченова Міністерства охорони здоров'я РФ спільно з Національною медичною бібліотекою США сформовано російсько-англійську версію тезауруса MeSH [29], який складається з двох частин: 1. Алфавітний перелік термінів з перехресними посиланнями (дескриптори (рубрики), модифікатори (підрубрики) та синоніми (перехресні посилання); 2. Ієрархічна структура, що має 11 рівнів. Перший рівень ієрархії включає 16 основних категорій. Основними лексичними одиницями MeSH є дескриптори і модифікатори. Кожен дескриптор має один або декілька унікальних номерів (кодів), що позначають його положення в ієрархічному дереві, яке може змінюватися по мірі оновлення MeSH.

Серед закордонних розробок також потрібно назвати комплекс тезаурусів Фонду Гетті (США, Каліфорнія), призначених для уніфікації описів об'єктів історико-культурної спадщини в найширшому сенсі цього слова, які широко застосовуються різними музеями, бібліотеками та архівами для опису власних зібрань. Для перегляду вони доступні в онлайн-режимі, проте повноцінне використання вимагає ліцензування.

На сторінці словників Getty Vocabularies представлені:

Тезаурус з мистецтва та архітектури (The Art and Architecture Thesaurus - AAT) – структуроване зібрання понять, що включає терміни з галузей мистецтва, архітектури, декоративного мистецтва, матеріальної культури та ін. У даний час тезаурус AAT містить близько 35 тисяч дескрипторів і понад 130 тисяч англомовних термінів. Дескриптори тезауруса поділяються на 7 фасетів: асоційовані поняття, фізичні властивості, стилі і періоди, агенти (люди і організації), діяльність, матеріали, об'єкти (Art and Architecture Thesaurus, 1994), а кожен фасет – на ієрархії. Всього налічується 33 ієрархічні рівні. Дескриптори тезауруса забезпечені стандартними для тезаурусів відношеннями вищий рівень-нижчий рівень і асоціативні зв'язки. Тезаурус доступний онлайн з вільною ліцензією[4].

Тезаурус географічних назв (The Getty Thesaurus of Geographic Names – TGN) [9], містить 1115000 дескрипторів, що включають назви держав, міст, археологічних об'єктів та їх фізичних особливостей, важливих для дослідження пам'яток мистецтва і архітектури.

Тезаурус імен авторів художніх творів (The Union List of Artist Names - ULAN) [32], що включає 375000 імен, біографічну і бібліографічну інформацію про художників і архітекторів, містить варіанти імен, псевдонімів і варіантів написання імен різними мовами.

Тезаурус «авторитетних файлів» – офіційних назв об'єктів історико-культурної спадщини, створений у 2011 році [18]. У тезаурусі зібрані офіційні назви та інші метадані творів мистецтва, архітектури, архівної документації, документів з фондів рідкісної книги, необхідні для здійснення їх уніфікованого опису. До складу словника включені авторитетні іконографічні поняття, складені спеціальною робочою групою Дослідницького інституту Гетті. Ця частина тезауруса включає власні імена персонажів (вигаданих, легендарних і реальних), а також найменування міфологічних, релігійних і літературних сюжетів, подій і місць, отриманих з найбільш відомих творів літератури, живопису та декоративно-прикладного мистецтва.

Тезауруси Дослідницького інституту Гетті узгоджені з міжнародними стандартами і можуть бути використані для підвищення ефективності пошуку в окремих базах даних та в Інтернеті. Вони можуть використовуватися при індексуванні архівних та бібліотечних ресурсів.

Міжнародною Продовольчою і сільськогосподарською організацією ООН – ФАО (Food and Agriculture Organization, FAO) розроблено тезаурус AGROVOC [1] сільськогосподарської тематики. Він охоплює термінологію сільсько-, лісо-, водо-, рибогосподарського спрямування, проблем механізації і будівни-

цтва, охорони природних ресурсів, запобігання забруднення навколишнього середовища та ін. Тезаурус містить понад 40 тис. концептів 29 мовами (в тому числі і українською). У ньому можна знайти, наприклад, назву будь-якої рослини мовою, що вас цікавить, а також семантичні зв'язки, що існують між сировинним товаром і сільськогосподарською культурою, з якої він виготовлений.

Сьогодні AGROVOC використовується науковцями, бібліотекарями та інформаційними менеджерами для індексування, пошуку та організації даних в сільськогосподарських інформаційних системах і на веб-сторінках.

Найбільшим у Росії центром наукової інформації і багатопрофільним науково-дослідним інститутом в області соціальних і гуманітарних наук Інститутом наукової інформації з питань суспільних наук Російської академії наук (ІНІСН РАН) розроблено комплекс галузевих тезаурусів з правознавства, економіки та демографії, політології, релігієзнавства, мовознавства, гендерних досліджень, соціології, які розміщено на сайті інституту http://inion.ru/resources/bazy-dannykh-inion-ran/ [15]

У 2018 році інститутом розроблено Інформаційно-пошуковий тезаурус ІНІСН з соціології [17], який включає 4760 термінологічних статей (з них 3126 дескрипторів і 1634 аскрипторів), що відображають різні аспекти соціологічної науки, і призначений для індексування вхідного потоку документів і запитів, забезпечуючи інтелектуальний пошук інформації в бібліографічному банку даних, в тому числі – пошук в режимі віддаленого доступу. Включає 3 покажчики: алфавітний лексико-семантичний, пермутаційний і систематичний.

Основною частиною тезауруса є алфавітний лексико-семантичний покажчик, де представлені всі дескриптори і аскриптори (синоніми) зі словниковими статтями. У тезаурусі використані недиференційовані ієрархічні відношення, асоціативні відношення (посилання на семантичні (не ієрархічні) зв'язки з іншими дескрипторами) та відношення синонімії заголовного дескриптора з аскриптором.

Пермутаційний покажчик є допоміжним до алфавітного лексико-семантичного. Містить дескриптори і аскриптори без словникових статей. Специфічною особливістю покажчика є формування «словникових гнізд» за ключовими словами, що входять до складу термінів-словосполучень.

Систематичний покажчик містить тільки дескриптори. Терміни розподілені в алфавітному порядку за рубриками ієрархічної класифікації рубрикатора ІНІСН. При цьому окремий дескриптор може бути включений в декілька рубрик.

У тому ж році інститутом розроблено Інформаційно-пошуковий тезаурус ІНІСН з етнології та антропології [16]. Містить 6698 термінів, що відображають різні аспекти етнології та антропології з дотриманням принципів історизму; включає два покажчика: алфавітний лексико-семантичний і пермутаційний.

Галузеві тезауруси ІНІСН входять до складу Великого інформаційного словника з суспільних наук (БІСОН), забезпечуючи як галузевий, так і міжгалузевої пошук інформації.

Ще одним відомим ресурсом є Тезаурус NASA [22], який містить авторизовані тематичні терміни Національного управління з аеронавтики і дослідження космічного простору (НАСА) (National Aeronautics and Space Administration), що використовуються для індексації та отримання матеріалів у Сервері технічних звітів НАСА. Сфера використання цього контрольованого словника не тільки аерокосмічна інженерія, але й усі галузі техніки та фізики, природні космічні науки (астрономія, астрофізика, планетарна наука), науки про Землю і біологічні науки. Тезаурус NASA містить понад 18 400 тематичних термінів, 4300 визначень та понад 4500 перехресних посилань.

В Україні робота із розроблення комп'ютерних термінологічних тезаурусів різних типів, як загальномовних, так і вузькогалузевих, проводиться співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка в рамках наукових студій з формалізації мовних досліджень.

Так науковцями розроблено Тезаурус з комп'ютерної лексикографії [26] (онлайн-версія розміщена на сторінках Лінгвістичного порталу MOVA.info у розділі «Словники» http://www.mova.info/Page.aspx?l1=61), що представляє терміносистему комп'ютерної ідеографії, що налічує лише 75 термінів.

Макроструктура тезауруса репрезентована родо-видовим деревом термінів. Нульовий рівень тезауруса представлено терміном *комп'ютерна лексикографія*, який є родовим для концепту першого рівня *комп'ютерна ідеографія*. Другий рівень має чотири концепти: одиниці КТ, відношення між одиницями КТ, комп'ютерний тезаурус (КТ) та укладання КТ, які містять відповідно 5, 8, 10 і 6 термінів третього рівня. Максимальна глибина ієрархізації Тезауруса з комп'ютерної лексикографії складає шість рівнів. Терміни розташовані в тематично-алфавітному прядку і об'єднуються видовими, родовими та синонімічними відношеннями. Словникова стаття тезауруса складається з заголовної одиниці-терміна та дефініції.

Співробітниками цієї ж установи розроблено тезаурус лінгвістичних термінів [27], який знаходиться у вільному доступі на сторінках Лінгвістичного порталу MOVA.info у розділі «Словники» http://www.mova.info/Page.aspx?l1=61 .

Тезаурус містить вживані загальнолінгвістичні терміни; терміни окремих прикладних напрямів лінгвістики; терміни з комп'ютерної лінгвістики, пов'язані з автоматизацією лінгвістичних процесів. Кількість термінів у тезаурусі – 3394 одиниці, які пов'язують 9265 семантичних відношень. Основними відношеннями є гіпонімія, парціація (частина – ціле), синонімія, кореляція, асоціація, локалізація об'єкта, його призначення, функція, відношення, імплікація тощо. Дескриптори мають перекладні аналоги англійською та російською мовами тоді як дефініція – українською.

Відповідь на запит користувач отримує у формі тезаурусного графа, зображеного у вигляді семантичної мережі – ієрархічно організованої структури даних (термінів-вузлів та дуг), які презентують різні типи відношень та подаються у тезаурусі в текстовому вигляді.

Активне впровадження в практичну роботу бібліотек автоматизованих технологій вимагають високого лінгвістичного забезпечення автоматизованих інформаційних бібліотечних систем (АІБС). Уніфікації пошукових термінів та ведення діалогу між людиною та комп'ютером з використанням інформаційно-пошукової мови сприяють саме тезауруси.

Цьому сприяє розробка фахівцями Національної бібліотеки України імені Ярослава Мудрого (Національної парламентської бібліотеки України), якими сформовано «Інформаційно-пошуковий тезаурус» Національної бібліотеки України імені Ярослава Мудрого, котрий призначеній для відображення змісту документів і запитів користувачів з метою формування технологій ефективного пошуку в автоматизованих інформаційно-бібліотечних системах [7].

Тезаурус Національної бібліотеки України імені Ярослава Мудрого може використовуватися в бібліотечних та інформаційних установах, де аналітико-синтетичне опрацювання документів здійснюється за допомогою уніфікованої пошукової мови, а також як термінологічний словник зі структурними зв'язками, що відображають місце терміна у системі понять. Будь-яка бібліотека або інша інформаційна установа, за допомогою цього тезауруса може формувати власні авторитетні файли предметних рубрик або визначати ключові слова. Використання уніфікованої інформаційно-пошукової мови, при якісному індексуванні документів, у майбутньому забезпечить високу точність пошуку в автоматизованих інформаційно-бібліотечних системах (АІБС) і дозволить оперативно реагувати на появу нових термінів.

Тезаурус містить 34 690 термінів, у тому числі 14,6 тис. дескрипторів, 5 тис. аскрипторів. У ньому наявні три класи основних відношень: еквівалентні («*В*» означає «вживається» – дескриптор, НВ означає «не вживати» – аскриптор), ієрархічні («Ш» – «ширший термін» – родовий термін, «Н» означає «вужчий термін» – видовий термін), асоціативні («А» – «асоціативний термін» – термін якимось чином пов'язаний з даним поняттям, але не є його синонімом, родовим або видовим поняттям).

До тезауруса включені універсальні терміни, що означають: дії і процеси; поняття наук; назви організацій і установ; види, типи і назви матеріалів, біологічних організмів, хімікатів, сільськогосподарських культур, порід тварин тощо; назви народів, мов (природних, штучних, комп'ютерних), небесних тіл; види мистецтва, літератури і архітектури; соціальні і природні явища (війни, революції, битви, катастрофи та ін.); назви священних книг, міфіч-

них персонажів; континенти і регіони; країни, їхній адміністративно-територіальний поділ (головним чином України); назви етнічних земель, елементів земної поверхні (океанів, морів, гір тощо), парків, заповідних зон.

Інформаційно-пошуковий тезаурус складається з 8 розділів: Лексико-семантичне зібрання термінів (загальний розділ); Покажчик географічних назв; Покажчик персоналій; Покажчик установ і організацій; Покажчик формальних підзаголовків; Покажчик російських відповідників; Покажчик термінів латинською мовою; Покажчик термінологічних джерел.

Проте до дескрипторів у тезаурусі подаються лише російські відповідники та відповідники латинською мовою на переважну більшість термінів з біології та медицини, а словникова стаття – тільки українською мовою.

Актуалізація тезауруса здійснюється саме власником цього інформаційного продукту, Національною бібліотекою України імені Ярослава Мудрого централізовано, з визначеною періодичністю.

Науковцями Харківський державний університет харчування та торгівлі (ХДУХТ) створено двомовний (українсько-російський) галузевий тезаурус лінгвістичного забезпечення електронної бібліотеки ХДУХТ [8]. Тезаурус складається з алфавітного і систематичного покажчика. Алфавітний покажчик містить алфавітний перелік дескрипторних статей. Дескрипторна стаття має заголовний дескриптор; ключові слова з класу еквівалентності; дескриптори, що підпорядковуються заголовному; дескриптори, асоційовані з заголовним. Систематичний покажчик служить для розкриття, обліку та контролю парадигматичних відношень між дескрипторами. Текстовий файл тезауруса становить собою контрольовану систему рубрик, пов'язаних між собою ієрархічними і асоціативними зв'язками. Галузевий тезаурус ХДУХТ містить терміни з розділів «Плоди та овочі» і «Рослини та пряноші». Він постійно доповнюється новими даними.

Проведений аналіз літератури показує, що маємо спроби розроблення вузькоспеціалізованих галузевих тезаурусів. Так, науковці Таврійського національного університету працюють над створенням тезауруса предметної області «Нанотехнології» [6], Одеського національного університету імені І.І. Мечникова – тезауруса будівельної термінології [28], Житомирського державного університету імені Івана Франка над побудовою двомовного (українсько-англійського) тезауруса авіації [5].

Висновки. Розглянуті у роботі інформаційно-пошукові тезауруси є необхідними допоміжними засобами орієнтації у потужних інформаційних потоках. Принципи побудови і функціонування розглянутих інформаційно-пошукових тезаурусів, дають можливість стверджувати, що специфіка предметної області кожного тезауруса знаходить відображення в його структурі.

Список використаних джерел

1. AGROVOC Multilingual Thesaurus. URL: http://agrovoc.uniroma2.it/agrovoc/agrovoc /en/index?clang=uk

2. Алексеев, А.А., Добров Б.В., Лукашевич Н.В. Лингвистическая онтология – тезаурус РуТез. Материалы 3 международной научно-технической конференции Открытые семантические технологии проектирования интеллектуальных систем. 2013. с.153-158.

3. Антоненко С. Актуальні питання створення та впровадження української версії тезаурусу «EUROVOC». Наукові записки Інституту законодавства Верховної Ради України. 2014. № 5. С. 79-84.

4. Art & Architecture Thesaurus. URL: http://www.getty.edu/research/tools/vocabularies /aat/index.html

5. Асмукович І. В. Принципи побудови двомовного тезауруса фахової мови авіації. Науковий вісник СНУ імені Лесі Українки : Філологічні науки : зб. наук. праць. Луцьк : СНУ. 2013. № 20 (269). С. 39-44.

6. Бержанский В., Нестеров Д. Разработка многоязычного словаря предметной области «Нанотехнологии». URL: http://megaling.ulif.org.ua/tezi-2012-rik/berzhanskiy-vladimir-naumovich-nesterov-dmitriy-sergeevich-razrabotka-mnogoyazyichnogo-slovarya-predmetnoy-oblasti-nanotehnologii.html

7. Вилегжаніна Т. І. Лінгвістичне забезпечення для створення уніфікованого пошукового образу та пошукового запиту документа. URL: http://library.kr.ua/conference/vylegzhanina.html

8. Губренко І. Ю. Створення галузевого тезауруса в лінгвістичному забезпеченні електронної бібліотеки Харківського державного університету харчування та тогівлі. Всеукраїнська наук.-практ. конф.: [до 20-річчя створення АСІБТ та 45-річчя заснування ХДУХТ : матеріали]. Харків : ХДУХТ. 2012. С. 74-78.

9. Getty Thesaurus of Geographic Names. URL: http://www.getty.edu/research/tools /vocabularies/tgn/index.html

10. GermaNet. URL: http://www.sfs.uni-tuebingen.de/GermaNet/

11. DanNet. URL: https://cst.ku.dk/projekter/dannet/

12. EuroVoc. URL: http: //europa. eu/eurovoc/

13. EuroVoc 4.9. URL: https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoci

14. EuroWordNet. URL: http://projects.illc.uva.nl/EuroWordNet/

15. ИНИОН РАН. URL: http://inion.ru/resources/bazy-dannykh-inion-ran/

16. Информационно-поисковый тезаурус Этнология. Антропология. URL: http://inion.ru/site / assets/files/2980/tezaurus_etnologiia.pdf

17. Информационно-поисковый тезаурус Социология. URL: http://inion.ru/site/assets /files/2641/tezaurus_sotciologiia_versiia_5-1.pdf

18. Cultural Objects Name Authority. URL: http://www.getty.edu/research/tools /vocabularies/cona/index.html

19. Кульчицький І. М., Романюк А. Б., Харів Х. Б. Розроблення Wordnet-подібного словника української мови. Вісник Національного університету «Львівська політехніка». 2010. № 673. С. 306-318.

20. MeSH. URL: https://www.ncbi.nlm.nih.gov/mesh

21. MultiWordNet. URL: http://multiwordnet.fbk.eu/english/home.php

22. NASA Thesaurus. URL: https://www.sti.nasa.gov/nasathesaurus

23. RussNet. URL: http://project.phil.spbu.ru/RussNet/index_ru.shtml

24. РуТез. URL: http://www.labinform.ru/pub/ruthes/index.htm

25. SNOMED. URL: http://www.snomed.org/snomed-ct/snomed-in-action

26. Тезаурус з комп'ютерної лексикографії. URL: http://www.mova.info/Page3. aspx?l1=188&vocid=1

27. Тезаурус з лінгвістичної термінології. URL: http://www.mova.info/mov_thes.aspx?l1=68

28. Філюк Л. М. Моделювання тезауруса української будівельної термінології. Мова. 2014. №14. С. 127-131.

29. Web-ИРБИС. URL: http://library.gpntb.ru/cgi-bin/irbis64r/62/cgiirbis_64.exe?C21COM=F&I21DBN=MESH&S21FMT=web_mesh_wn&S21All=%3C.%3ER=0%3C.%3E&P21DBN=IBIS&Z21ID

30. WordNet. URL: https://wordnet.princeton.edu/

31. UNESCO Thesaurus. URL: http://skos.um.es/unescothes/

32. Union List of Artist Names. URL: http://www.getty.edu/research/tools/vocabularies /ulan/index.html

References

1. AGROVOC Multilingual Thesaurus, Retrieved from: http://agrovoc.uniroma2.it/agrovoc/agrovoc/en/index?clang=uk [in English].

2. Alekseev, A.A., Dobrov B.V., Lukashevich N.V. (2013). Lingvisticheskaya ontologiya – tezaurus RuTez. Materialyi 3 mezhdunarodnoy nauchno-tehnicheskoy konferentsii Otkryityie semanticheskie tehnologii proektirovaniya intellektualnyih sistem, pp.153-158. [in Russian]

3. Antonenko S. (2014). AktualnI pitannya stvorennya ta vprovadzhennya ukraYinskoYi versIYi tezaurusu «EUROVOC». NaukovI zapiski Institutu zakonodavstva VerhovnoYi Radi UkraYini, # 5, pp. 79-84. [in Ukrainian]

4. Art & Architecture Thesaurus, Retrieved from: http://www.getty.edu/research/tools/vocabularies/aat/index.html. [in English].

5. Asmukovich I. V. (2013). Printsipi pobudovi dvomovnogo tezaurusa fahovoYi movi avIatsIYi. Naukoviy vIsnik SNU ImenI LesI UkraYinki : FIlologIchnI nauki : zb. nauk. prats. Lutsk: SNU, # 20 (269), pp. 39-44. [in Ukrainian]

6. Berzhanskiy V., Nesterov D. Razrabotka mnogoyazyichnogo slovarya predmetnoy oblasti «Nanotehnologii». Retrieved from: http://megaling.ulif.org.ua/tezi-2012-rik/berzhanskiy-vladimir-naumovich-

nesterov-dmitriy-sergeevich-razrabotka-mnogoyazyichnogo-slovarya-predmetnoy-oblasti-nanotehnologii.html [in Ukrainian]

7. VilegzhanIna T. I. LIngvIstichne zabezpechennya dlya stvorennya unIfIkovanogo poshukovogo obrazu ta poshukovogo zapitu dokumenta. Retrieved from: http://library.kr.ua/conference/vylegzhanina.html [in Ukrainian]

8. Gubrenko I. Yu. (2012). Stvorennya galuzevogo tezaurusa v lIngvIstichnomu zabezpechennI elektronnoYi bIblIoteki HarkIvskogo derzhavnogo unIversitetu harchuvannya ta togIvlI. VseukraYinska nauk.-prakt. konf.: [do 20-rIchchya stvorennya ASIBT ta 45-rIchchya zasnuvannya HDUHT : materIali]. HarkIv : HDUHT, pp. 74-78 [in Ukrainian]

9. Getty Thesaurus of Geographic Names, Retrieved from: http://www.getty.edu /research/tools/vocabularies/tgn/index.html. [in English].

10. GermaNet, Retrieved from: http://www.sfs.uni-tuebingen.de/GermaNet/ [in German].

11. DanNet, Retrieved from: https://cst.ku.dk/projekter/dannet/

12. EuroVoc, Retrieved from: http: //europa. eu/eurovoc/ [in English].

13. EuroVoc 4.9, Retrieved from: https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoси. [in English].

14. EuroWordNet, Retrieved from: http://projects.illc.uva.nl/EuroWordNet/. [in English].

15. INION RAN, Retrieved from: http://inion.ru/resources/bazy-dannykh-inion-ran/ . [in Russian]

16. Informatsionno-poiskovyiy tezaurus Etnologiya. Antropologiya, Retrieved from: http://inion.ru/site/assets/files/2980/tezaurus_etnologiia.pdf. . [in Russian]

17. Informatsionno-poiskovyiy tezaurus Sotsiologiya, Retrieved from: http://inion.ru/site /assets/files /2641/tezaurus_sotciologiia_versiia_5-1.pdf. . [in Russian]

18. Cultural Objects Name Authority, Retrieved from: http://www.getty.edu/research /tools/vocabularies/cona/index.html.[in English].

19. Kulchitskiy I. M., Romanyuk A. B., HarIv H. B. (2010). Rozroblennya Wordnet-podIbnogo slovnika ukraYinskoYi movi. VIsnik NatsIonalnogo unIversitetu «LvIvska polItehnIka»,# 673, pp. 306-318. [in Ukrainian]

20. MeSH, Retrieved from: https://www.ncbi.nlm.nih.gov/mesh. [in English].

21. MultiWordNet, Retrieved from: http://multiwordnet.fbk.eu/english/home.php. [in Italian].

22. NASA Thesaurus, Retrieved from: https://www.sti.nasa.gov/nasathesaurus. [in English].

23. RussNet, Retrieved from: http://project.phil.spbu.ru/RussNet/index_ru.shtml. [in Russian]

24. РуТез, Retrieved from: http://www.labinform.ru/pub/ruthes/index.htm. [in Russian]

25. SNOMED, Retrieved from: http://www.snomed.org/snomed-ct/snomed-in-action. [in English].

26. Tezaurus z komp'yuternoYi leksikografIYi, Retrieved from: http://www.mova.info/Page3.aspx?l1=188&vocid=1. [in Ukrainian]

27. Tezaurus z lIngvIstichnoYi termInologIYi, Retrieved from: http://www.mova.info/mov_ thes.aspx?l1=68 [in Ukrainian]

28. FIlyuk L. M. (2014). Modelyuvannya tezaurusa ukraYinskoYi budIvelnoYi termInologIYi. Mova, #14. pp. 127-131. [in Ukrainian]

29. Web-ИРБИС, Retrieved from: http://library.gpntb.ru/cgi-bin/irbis64r/62/cgiirbis_64.exe?C21COM=F&I21DBN=MESH&S21FMT=web_mesh_wn&S21All=%3C.%3ER=0%3C.%3E&P21DBN=IBIS&Z21ID. . [in Russian]

30. WordNet, Retrieved from: https://wordnet.princeton.edu/.[in English].

31. UNESCO Thesaurus, Retrieved from: http://skos.um.es/unescothes/.[in English].

32. Union List of Artist Names, Retrieved from: http://www.getty.edu/research /tools/vocabularies/ulan/index.html. [in English].