УДК 004.512

О. В. БІСІКАЛО, О. В. ЯХИМОВИЧ

АВТОМАТИЗОВАНЕ ВИЗНАЧЕННЯ ЛЕКСИЧНИХ ОНТОЛОГІЙ З ТЕЗАУРУСУ ТЕХНІЧНОГО СПРЯМУВАННЯ

Вінницький національний технічний університет,
21021, вул. Хмельницьке шосе 95, м. Вінниця, Україна,
E-mail: obisikalo@vntu.edu.ua

Анотація. Запропоновано підхід до автоматизованого визначення лексичних технологій технічного спрямування. Задані специфікація та вимоги до побудови тезаурусу, визначені можливі типи відношень між термінами у тезаурусі. Побудувано та протестовано програмну процедуру для побудови онтологій з синонімічних ланцюгів на основі тлумачного словника.

Ключові слова: Тезаурус, автоматизація, лексична онтологія, специфікація, відношення, синонімія, програмна процедура, NET Framework.

Аннотация. Предложен подход к автоматизированному определению лексических технологий технического направления. Заданы спецификация и требования к построению тезауруса, определены возможные типы отношений между терминами в тезаурусе. Построено и протестировано программную процедуру для построения онтологий из синонимических цепей на основе толкового словаря.

Ключевые слова: Тезаурус, автоматизация, лексическая онтология, спецификация, отношение, синонимия, программная процедура, NET Framework.

Abstract. The approach to the automated definition of lexical technologies of a technical direction is offered. Specification and requirements for building a thesaurus are given, possible types of relations between terms in the thesaurus are defined. A program procedure for constructing ontologies from synonymous chains based on an explanatory dictionary has been constructed and tested.

Keywords: Thesaurus, automation, lexical ontology, specification, video, synonym, program procedure, NET Framework.

ВСТУП

У час бурхливого розвитку інформаційного суспільства актуалізувалась проблема застосування адекватних формальних моделей у лінгвістиці, зокрема — у лексикографії. Саме вони уможливлюють створення сучасних інтелектуальних лінгвотехнологій у природомовних людино-машинних системах [1].

На сьогодні одним з найважливіших завдань лексикографії є проектування таких словників, які б на рівні світових стандартів задовольняли велику потребу сучасної інформатизованої спільноти в систематизованій лінгвістичній інформації. З огляду на це тезауруси як словники, які не лише інвентаризують, а й систематизують лексичні одиниці у межах певної мовної підсистеми, потрапляють у поле підвищеної уваги фахівців. Рівень розвитку інформаційних технологій в Україні дозволяє, а потреби користувача вимагають зосередитися на розробленні тезаурусів різних типів: як загальномовних, так і вузькогалузевих термінологічних. Історія укладання ідеографічних словників має довгу традицію: однією з найдавніших писемних пам'яток тезаурусного типу є створений ще в II-III століттях до н. е. санскритський словник «Амара-коша». Серед наукових праць, актуальних і на сьогодні, найбільш відомі такі тезауруси: словник П. Роже для англійської мови, П. Буассьєра — для французької, Ф. Дорнзайфа — для німецької, Х. Касареса — для іспанської. Вагомий внесок у розбудову тезаурусної і дотичної до тезаурусної проблематики зробили Ш. Баллі, Л. В. Щерба, Н. Ю. Шведова, В. В. Морковкін, Ю. М. Караулов, Ю. Д. Апресян, О. С. Баранов, І. О. Мельчук, М. А. Кронгауз, М. Я. Гловінська,

Л. Г. Бабенко. Серед українських дослідників можна відзначити роботи таких авторів, як В. А. Широков, Н. П. Дарчук, В. В. Дубічинський, І. М. Гетьман, Н. В. Сніжко, М. Д. Сніжко, А. Я. Середницька.

У комп'ютерній мережі Інтернет успішно функціонують і розвиваються Прінстонський тезаурус WordNet, MindNet — програмний продукт проекту дослідницького відділу Майкрософт, FrameNet, розроблений в університеті Берклі, VerbNet, HowNet, ConceptNet, багатомовна електронна лінгвістична база EuroWordNet, створені за аналогією до них бази GermaNet, BalkaNet, RusNet та багато інших.

Кінцевим продуктом розробки є саме комп'ютерний тезаурус як найоптимальніша за впорядкуванням лінгвістичного матеріалу та швидкодією система, а тому потребує удосконалення формалізована, алгоритмізована методика його побудови.

Тезаурус — це термін, який широко використовується в галузі інформатики як складова частина інформаційно-пошукових систем. Можна розглядати два визначення тезауруса:

— тезаурус — це словник, що відображає семантичні відношення між лексичними одиницями дескрипторної інформаційно-пошукової мови (дескрипторами) і призначений для пошуку слів за їх смисловим змістом;

— тезаурус — це контрольований словник термінів предметної області, який створюється для поліпшення якості інформаційного пошуку в певній предметній сфері.

І в тому, і в іншому випадку мова йде про словник, що має полегшити пошук необхідної інформації. У словниках можливі два способи розташування слів: за близькістю їх літерного складу та за смисловою близькістю.

Словник — це абстрактний мовно-інформаційний об'єкт, визначальною рисою якого передусім є членоване розміщення матеріалу — основною композиційною та комунікативною одиницею слугує відносно самостійний відрізок тексту, який називають словниковою статтею [2].

Тезауруси створюються саме за смисловою близькістю. Якщо словники, скоріше, розкривають значення будь-якого слова, то тезауруси створюються таким чином, щоб пошукове слово виражало певне якесь поняття на основі закладених в ньому взаємозв'язків.

Тезаурус є невід'ємною частиною пошукової системи і являє собою ієрархічну мережу понять, що відповідають тим чи іншим значенням окремих слів або текстових виразів.

Серед найбільш перспективних напрямів розвитку автоматичних тезаурусів необхідно визначити такі:

— одержання довідки за пошуковим словом: вказавши слово як джерело запиту, користувач у відповідь отримує потрібний фрагмент словника, який вміщує лінгвістичну інформацію про це слово;

— контекстні заміни на вимогу користувачів: у цьому випадку тезаурус підбирає замість одного словосполучення інше, котре користувач визнає як таке, що більш відповідає контексту за смислом або стилем [3].

У словнику-тезаурусі мовні вирази формують семантичні поля. Семантичне поле визначається як область дійсності, що має в мові відповідність у вигляді тематично об'єднаної сукупності мовних (переважно лексичних) одиниць. Семантичне поле характеризується зв'язком значень мовних одиниць, системним характером цих зв'язків, взаємозалежністю і взаємовизначеністю мовних одиниць, відносною автономність поля, безперервністю смислового простору, видимістю і психологічною реальністю для середнього носія мови. Між одиницями семантичного поля можуть встановлюватися синонімічні та антонімічні, гіперо-гіпонімічні, логічні (що відображають логіку пізнання людиною світу) і асоціативні відносини. Відзначаючи цей факт. Ю. С. Степанов називає такі відносини «структурними лініями», які пронизують систему лексики і орієнтуючим людини в пошуку необхідної інформації.

При спільності змісту одиниці поля можуть мати формальні відмінності, в залежності від яких виділяються функціонально-семантичні поля, що включають різнорівневі засоби мови (лексичні, морфологічні, синтаксичні), граматичні поля — об'єднання граматичних засобів вираження деякого граматичного значення [4], словотворчі , або морфосемантичні поля — об'єднання слів, семантична близькість яких обумовлена наявністю загального афікса чи основи, а також лексичні поля, що включають слова, значення яких підведене під деяке загальне поняття [5]. Лексичні поля мають свої типи в залежності від природи понятійної суті, яка лежить в їх основі. Слова, що відносяться до однієї і тієї ж частини мови і мають спільність лексичних значень, формують лексико-семантичне поле [6, 7].

Більшість тезаурусів виділяють у складі певної мови невеликі за обсягом поля — лексико-семантичні групи, — одиниці яких є синонімами і антонімами. Самі ж групи залишаються семантично не пов'язаними між собою, вони шикуються в словнику за алфавітним принципом з урахуванням головного слова в групі.

Тезауруси, які пропонують аранжування широкого семантичного простору, що покривається всім словником даної мови, залишаються (в силу складності їх складання) явищем досить рідкісним. Організація такого семантичного простору відноситься до «логічного» типу — вона відображає логіку пізнання й категоризації світу людиною. При цьому понятійні категорії, які формують семантичний

простір лексикону і стають основою для лексичних полів, співвіднесені ієрархічно, тобто включаються один в одного на основі гіпер-гіпонімічних і/або партонімічних відносин. Серед словників логічного типу одним з найбільш авторитетних є тезаурус П. М. Роже, вперше опублікований в 1852 році. У ньому весь лексикон англійської мови спочатку співвіднесений з шістьма понятійними сферами: (1) АБСТРАКТНІ ВІДНОСИНИ, (2) ПРОСТІР, (3) РЕЧОВИНА, (4) ІНТЕЛЕКТ, (5) ВОЛЯ, (6) ПСИХОЛОГІЧНІ СТАНИ (афект-AFFECTIONS). Усередині кожної з цих сфер має місце своя детальна класифікація, в результаті якої виділяються 1000 семантичних категорій і 8 рівнів, співвіднесених за принципом ієрархічного включення.

В іншому відомому словнику — тезаурус Р. Халліга і В. Вартбурга — словниковий склад німецької мови стратифікований на такі понятійні класи і підкласи: ВСЕСВІТ — Небо і атмосфера. Земля. Рослинний світ. Тваринний світ. ЛЮДИНА — Людина як жива істота. Душа і розум. Людина як суспільна істота. Соціальна організація і соціальні інститути. ВСЕСВІТ І ЛЮДИНА — Наука і техніка. Апріорні категорії.

На другому рівні членування кожен з підкласів має свою понятійну стратифікацію. Наприклад, ЛЮДИНА: Людина як жива істота — (1) Пол. (2) Раса. (3) Частини тіла. (4) Органи та їх функціонування. (5) П'ять почуттів, (б) Рухи і положення тіла. (7) Сон. (8) Здоров'я і хвороби. (9) Людське життя взагалі. (10) Потреби людини як живої істоти. ДУША І РОЗУМ - (1) Загальні положення, розум, мудрість, здібності. (2) Сприйняття. (3) Свідомість, уявлення. (4) Пам’ять. (5) Уява. (6) Мислення. (7) Почуття. (8) Воля. (9) Мораль. Людини як суспільної істоти — (1) Суспільне життя взагалі: а) устрій суспільства; б) мова; в) суспільні зв'язки. (2) Людина в праці: а) загальні положення; б) сільське господарство; в) ремесла і професії; г) промисловість; д) торгівля; е) власність; ж) будинок. кімната. (3) Транспорт. (4) Пошта, телеграф, телефон. СОЦІАЛЬНА ОРГАНІЗАЦІЯ ТА СОЦІАЛЬНІ ІНСТИТУТИ - (1) Громадський колектив. (2) Держава. (3) Право. (4) Освіта, (б) Зовнішня політика. (6) Національна оборона. (7) Війна. (8) Література і мистецтво. (9) Віросповідання і релігія.

Сучасною версією тезауруса, організованого за «логічного» принципом, є словник-тезаурус російської мови О. С. Баранова. Організація цього словника подібна до тієї, що представлена в класичній роботі Роже. Слова, зібрані в гнізда (семи), групуються навколо певного поняття (ідеї) і, як правило, пов'язані родовидовими, або гіпер-гіпонімічними відносинами. В даний час словник включає 5923 гнізд і 7 рівнів ділення. Верхній рівень, представлений шістьма групами — (1) ПОРЯДОК, (2) ПРИРОДА, (3) ЛЮДИНА, (4) ДІЯЛЬНІСТЬ, (5) ТОВАРИСТВО, (б) КУЛЬТУРА — підрозділяється на 22 підгрупи, які, в свою чергу, поділяються на 76 відділів і т. д.

У тезаурусі логічного типу стратифікація даних порівнянна з полями Й. Тріра, який розділив весь словник на поля вищого рангу, далі поділені на поля більш низького рангу, і так до тих пір, поки не буде досягнутий рівень окремих, позначених словами, понять.

В іншому типі тезаурусів семантичний простір лексикону організовано по асоціативному принципом: слова групуються навколо слова-стимулу певної групи слів-асоціатів; такі у різних інформантів виявляють значний ступінь спільності. Психологічні асоціації встановлюються між предметами / поняттями, з одного боку, і між відповідними мовними одиницями, з іншого. Асоціації забезпечують «тяжіння» мовних одиниць один до одного. Прикладом такого тезауруса є словник сучасної російської мови, в основі якого лежить асоціативно-вербальна мережа. Пояснюючи принцип організації словника, Ю. М. Караулов відзначає, що ця мережа приводиться до вираженню в результаті анкетування великої кількості носіїв мови, спонтанно реагують на пропонований стимул, і є підстави говорити про її безпосередній зв'язок з мовною компетенцією. У мережі кожне слово присутнє у всьому різноманітті своїх значень, своїх синтаксичних і семантичних зв'язків з іншими словами, входячи в різні асоціативні поля.

Асоціативні поля подібні з синтагматичними полями, відомими також як поля Порцига. Вони утворені синтаксичними одиницями з семантично сумісними компонентами; наприклад, йти — ноги, гавкати — собака.

В цілому ж в тезаурусах різних типів стратифікація лексики спирається на інтуїцію і «наївну логіку», керуючись якою ми упорядковуємо знання про спостережуваному нами світі. Відповідно до цієї логіки визначаються рівні понятійних категорій і встановлюються гіпер-гіпонімічі і партонімічні відносини як структурні принципи організації конституентів цих категорій. Тим часом дані, представлені у словниках-тезаурусах (особливо в словниках асоціативного типу) свідчать про те, що крім гіперонімії і партонімії існують і інші відносини, які структурують понятійні простори і які слід враховувати при складанні словників. Тобто нам необхідна методологія, яка спирається не тільки на «наївну логіку» і інтуїцію носіїв мови, але й на деякі досить чіткі алгоритми, що застосовуються при побудові концептуальних моделей, або онтологій, словників-тезаурусів [8, 9].

Розробка електронних лінгвістичних словників в Україні становить надзвичайно актуальну проблему. Диктує умови інтенсивного розвитку не лише власне наукова потреба об’єктивації досліджень

(оптимізація та раціоналізація професійної роботи мовознавця), а й позанаукові аспекти — необхідно задовольнити зростаючий попит пересічного користувача на адаптовану для нього об'єктивну, достовірну інформацію мовознавчого характеру у вигляді електронних словників різного типу.

Автоматизовані лексикографічні бази у вигляді електронних словників зараз становлять невід'ємну частин систем машинного перекладу, інформаційного пошуку, редагування та правки текстів, а також обробки великих текстових масивів та їх зберігання як окремої задачі створення електронних бібліотек. Комп'ютерні словники на оптичних носіях дали змогу перекладачам, науковцям швидко знаходити будь-яку інформацію про слово таку, як переклад, орфографічну, граматичну інформацію, тлумачення тощо.

Створення тезаурусів, словників термінів дотепер залишається вкрай складною й трудомісткою роботою, ступінь автоматизації якої дуже низький. По суті, всі тезауруси створюють вручну. Автоматично може перевірятися лише узгодженість накопичених визначень. Альтернативою міг би бути підхід, коли визначення понять створюють за наявними текстами (енциклопедії, підручники, довідники), а потім, у разі необхідності, корегують в процесі діалогу з експертом.

Розробка тезаурусу для певної мови — проект яким займається команда щонайменше п'яти лінгвістів, зокрема лексикографів, упродовж трьох років за умови відповідного фінансування [10]. Тому автоматизація процесів вилучення знань з нього для прикладних лінгвістичних задач дасть змогу зменшити як кількість затрачених коштів, так і час, витрачений експертами-лінгвістами. Наприклад, автоматизована побудова онтологій на основі знаходження відношень між термінами, а саме відношень синонімії є актуальною проблемою, оскільки у більшості існуючих систем такі відношення формуються вручну. Ланцюги синонімів [2] можна використовувати для створення систем, які могли б шукати в тексті потрібний термін за його описом з метою реалізації феномену «*tip-of-the-tongue*» [11], замінювати слова на потрібні синоніми, а також знаходити різного ступеню прихованості варіанти плагіату в тексті.

ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Розглянемо тезаурус як повний систематизований набір даних про технічну галузь знань, що відповідає такій формальній моделі:

$$Ts = (T, R),$$

де T — скінчена множина термінів; R — скінчена множина відношень між цими термінами.

Тезаурус також можна розглядати як семантичну мережу (онтологію), у вузлах якої знаходяться терміни з T, що пов'язані відношеннями з обмеженого набору R.

Необхідно знайти підмножину термінів $T' \subseteq T$: $Ts' = (T', R') | Ts' \subseteq Ts, R' \subseteq R$, де Ts' — часткова онтологія з T, що відповідає типу відношень R' як підмножині відношень R між термінами.

Отже, за умови визначення R' як відношення синонімії, **мета** дослідження полягає в отриманні таких специфікації та вимог до побудови тезаурусу, які б дозволили визначити формальну процедуру для побудови онтологій з синонімічних ланцюгів на основі даних тезаурусу.

Для реалізації поставленої мети потрібно розв'язати такі задачі:

— задати специфікацію та вимоги до побудови тезаурусу;

— визначити типи відношень між термінами у тезаурусі;

— побудувати та протестувати програмну процедуру для побудови онтологій з синонімічних ланцюгів на основі даних тезаурусу.

1. СПЕЦИФІКАЦІЯ ТА ВИМОГИ ДО ПОБУДОВИ ТЕЗАУРУСУ

Будемо вважать основними технологічними фазами створення тезауруса такі:

— виділення лексичних одиниць, тобто формування словника (глосарія) T;

— розробка набору семантичних зв'язків;

— актуалізація зв'язків — установлення зв'язків між термінами.

При цьому дуже важливо сформулювати принципи, за якими буде здійснюватися кожна процедура. Для першого пункту визначальними є два аспекти — джерело лексичних одиниць та критерій їх добування. При розробці набору семантичних відношень можна знаходити їх у тексті, що описує дану галузь (намагатися вичленувати й уніфікувати ті відношення, що існують в текстах між термінами) або безпосередньо аналізувати знання. На практиці звичайно використовують поєднання обох методик. Для актуалізації семантичних зв'язків між термінами тезауруса можна використовувати знання експертів, а також документи, призначені як для фіксації структури знань (словники, класифікатори тощо), так і самі знання, що відображаються в рефератах, статтях, монографіях тощо.

Термін — це слово або словесний комплекс, що співвідноситься з поняттям визначеної організованої галузі пізнання (науки, техніки) та вступає у системні відношення з іншими словами,

словесними комплексами й утворює разом з ними в будь-якому окремому випадку чи у певний час замкнуту систему, що відрізняється високою інформативністю, однозначністю, точністю й експресивною нейтральністю.

Для створення тезауруса можна скористатися методологією розробки онтологічних моделей, згідно з якою побудова тезауруса складається з п’яти основних дій:

а) вивчення і систематизація початкових умов — мети і контексту розробки тезауруса;

б) збирання і накопичення даних;

в) аналіз даних;

г) початкова розробка тезауруса — встановлення зв'язків між термінами;

д) уточнення та затвердження тезауруса — аналіз користувачем отриманого тезауруса та його коректування.

Під час формування тезауруса доцільно враховувати наступні рекомендації, які стосуються для побудови визначень даних і метаданих та враховують вимоги, розроблені підкомітетом зі стандартизації ПК-6 «Телекомунікації та обмін інформацією між системами» з урахуванням ISO/IEC 11179.

Визначення термінів тезауруса може здійснюватися в автоматичному режимі (шляхом аналізу повнотекстових документів та інших інформаційних джерел), шляхом вилучення з інших баз знань (тезаурусів, онтологій тощо) або надаватися безпосередньо експертами.

Пропонуються такі формальні вимоги до визначень термінів у тезаурусі:

— визначення має бути викладене в однині. Виняток становлять поняття, які самі є множинними. Наприклад, «номер статті»: добре визначення — «номер посилання, що ідентифікує статтю»; погане визначення — «номер посилання для ідентифікації статей». У поганому визначенні використовується слово «статті», що може бути формою множини і це можна зрозуміти так, ніби один номер може посилатися на кілька статей;

— визначення повинне пояснювати, чим є наведене поняття, а не тільки чим воно не є. Наприклад, «розмір вартості фрахтування»: добре визначення — «розмір витрат, які несе вантажовідправник для переміщення товарів з одного місця до іншого»; погане визначення — «розмір витрат, що не належать до витрат на пакування, документальне оформлення, завантаження, розвантаження та страхування». У поганому прикладі не вказано, що входить до поняття елемента даних;

— визначення повинне мати вигляд описової фрази або речення. Речення необхідне для формування точного визначення, яке містить важливі характеристики поняття. Просте наведення одного або кількох синонімів не є достатнім. Наприклад, «ім’я агента»: добре визначення – «назва сторони, яка уповноважена діяти від імені іншої сторони»; погане визначення — «представник». «Представник» є синонімом імені елемента даних, який не може бути адекватним визначенням;

— визначення повинне містити лише широко відомі скорочення. Розуміння значення скорочення, зокрема абревіатур та ініціалів, зазвичай обмежується певним середовищем. В іншому середовищі ті ж самі скорочення можуть викликати неправильне розуміння або непорозуміння. Таким чином, для запобігання неоднозначності, у визначеннях використовуються тільки повні слова без скорочень. Наприклад, «прилад для вимірювання щільності»: добре визначення — «прилад, який використовується для вимірювання концентрації рідини, в одиницях виміру маси до одиниці об’єму (м. д. о.) (тобто фунтів на кубічний фут; кілограмів на кубічний метр)»; погане визначення «прилад, який використовується для вимірювання концентрації рідини в термінах м. д. о. (тобто фунтів на кубічний фут; кілограмів на кубічний метр)». Проте м. д. о не є загальновідомим скороченням і його значення може бути незрозумілим для деяких користувачів. Скорочення має бути наведене повними словами;

— визначення має бути викладене без використання визначень інших даних або базових понять. Визначення термінів має наводитись у відповідному глосарії. Якщо потрібне інше визначення, воно має додаватись як примітка після тексту первинного визначення або як окремий запис у словнику. Пов’язані визначення можна отримати за допомогою атрибутів посилання (перехресних посилань). Наприклад: «код типу зразка»: добре визначення — «код, який ідентифікує тип зразка»; погане визначення — «код, який ідентифікує тип обраного зразка. Зразок — це мала частка, вилучена для проведення експериментів. Він може бути як єдиним зразком для тестування, так і сурогатним зразком для контролю якості. Зразок для контролю якості — це сурогатний зразок, обраний для перевірки результатів тестування єдиних зразків». Погане визначення містить два додаткових визначення —«зразка» та «зразка для контролю якості».

Семантичними вимогами до визначень термінів у тезаурусі варто вважати:

— визначення має відображати суттєвий зміст поняття. Усі первинні характеристики поняття, мають бути відображені у визначенні з відповідним рівнем специфічності залежно від контексту. При цьому необхідно запобігати пояснення неважливих параметрів. Рівень деталізації залежить від потреб користувача системи та середовища.

Наприклад, «номер послідовності завантаження вантажу» (визначений контекст: будь-яка форма транспортування): добре визначення — «номер, що вказує на послідовність, в якій здійснюється завантаження до транспортного засобу або елемента транспортного середовища»; погане визначення — «номер, який відображає послідовність, в якій здійснюється завантаження до вантажівки» (у визначеному контексті вантажі можуть транспортуватись різними транспортними засобами:, вантажівками, кораблями, вантажними потягами і не обмежене лише вантажівками).

Інший приклад: «сума за рахунком-фактурою»: добре визначення — «загальна сума, яку потрібно сплатити за рахунком-фактурою»; погане визначення — «загальна сума вартості всіх елементів, зазначених в рахунку-фактурі, включаючи усі відрахування, зокрема знижки та дисконти, додаткові платежі, зокрема страхові, транспортні та накладні витрати тощо». У поганому визначенні міститься зайва інформація;

— визначення має бути точним та однозначним, достатньо зрозумілим, щоб забезпечити його однозначну інтерпретацію. Наприклад, «дата отримання вантажу»: добре визначення — «дата, на яку вантаж передається отримувачу»; погане визначення — «дата, на яку здійснюється доставка вантажу». У поганому визначенні не роз'яснюється, що таке «доставка». Під «доставкою» можна зрозуміти як момент розвантаження товару у певному місті, так і факт передачі товару кінцевому отримувачу. Не виключено, що кінцевий отримувач ніколи не отримає вантаж або його передача може здійснитися через кілька днів після розвантаження;

— визначення має бути коротким. Слід запобігати використання додаткових фраз описового характеру, подібних до «для забезпечення використання цього реєстру метаданих», «терміни, що мають бути описані». Наприклад, «ім'я набору символів»: — добре визначення «ім'я, що присвоюється набору фонетичних або ідеографічних символів, в які зашифровані дані»; погане визначення — «м'я, що присвоюється набору фонетичних або ідеографічних символів, в яких зашифровані дані для забезпечення використання цього реєстру метаданих або, якщо говорити про загальний вжиток, спроможність системного обладнання і програмного забезпечення обробляти дані, зашифровані одним або декількома шифрами». У поганому визначенні всі фрази після виразу «… в яких зашифровані дані» є зайвими;

— визначення повинне мати можливість використовуватися окремо. Зміст поняття має бути наочним у визначенні. Для розуміння поняття не потрібні додаткові роз'яснення. Наприклад, «назва міста розміщення школи»: добре визначення — «назва міста, де знаходиться школа»; погане визначення — «див. сайт школи». Погане визначення не є самостійним, оскільки необхідно звернутися до додаткового джерела;

— визначення повинне бути поданим без використання пояснювальної інформації, функціонального використання або процедурної інформації. Пояснення не слід включати до визначень, тому що вони містять зайву інформацію. У разі потреби такі пояснення можуть бути розміщені в інших атрибутах метаданих. Припустимо додати кілька прикладів після визначення. Наприклад, «мітка поля даних»: добре визначення — «ідентифікація поля в індексі, тезаурусі, базі даних тощо»; погане визначення — «ідентифікація поля в індексі, тезаурусі, базі даних тощо, яка застосовується для таких елементів інформації як примітки, колонки в таблицях». У поганому визначенні містяться примітки, що стосуються функціонального використання. Якщо інформація, що починається зі слів «яка застосовується…» є необхідною, то вона має бути розміщена в іншому атрибуті;

— визначення повинне запобігати циклічних посилань. Два поняття не слід розкривати одне через одне. Визначення одного поняття не може використовувати інше поняття як своє визначення, тому що це може призвести до ситуації, коли поняття визначається через інше поняття, яке, у свою чергу, визначається через перше поняття. Наприклад, два елементи даних з поганими визначеннями — «ідентифікаційний номер працівника — номер, що призначається працівнику; «працівник — людина, яка має відповідний ідентифікаційний номер працівника». Визначення посилаються одне на одне, але в жодному з них не наведено зміст поняття;

Визначення повинні використовувати однакову термінологію та логічну структуру для пов'язаних визначень. Для близьких або пов'язаних визначень має використовуватись одна й та ж сама термінологія та синтаксис. Наприклад, «дата відправлення товарів — дата, в яку товари були відправлені даній стороні», «дата отримання товарів — дата, в яку товари були отримані даною стороною». Використання єдиної термінології значно спрощує розуміння [10].

2. ВИЗНАЧЕННЯ ВІДНОШЕНЬ МІЖ ТЕРМІНАМИ В ТЕЗАУРУСІ

Основним відношенням (зв`язком) між термінами в тезаурусі є зв`язок між ширшими (виразнішими) і вужчими (більш спеціалізованими) поняттями. Виділяють два підвиди цього відношення:

— один термін позначає поняття, що є частиною поняття, що позначається іншим терміном (наприклад, «видавництво» і «друкарня»);

— один термін позначає елемент класу, що позначається іншим терміном («спеціальні види друку» і «райдужний друк»).

Це відношення на множині термінів є відношенням часткового порядку, тобто множина термінів з такими зв`язками утворює ациклічний граф, або поліієрархічну структуру.

Існують також і інші зв`язки між термінами. Наприклад, одне поняття або концепцію може бути позначено декількома термінами, синонімами. Деякі терміни можуть бути антонімами для інших. Часто серед термінів, що відносяться до одного поняття, виділяють єдиний (для кожної мови тезауруса), найбільш переважний (найбільш відповідний) термін, який найкраще характеризує або позначає дане поняття. Решта термінів є менш переважними (менш відповідними).

Окрім вищеописаних, між термінами можуть існувати також інші, загалом асоціативні зв'язки, якщо поняття, що позначаються цими термінами, як-небудь пов`язані між собою за своїм сенсом, за винятком описаних вище ієрархічних зв'язків.

У багатомовних тезаурусах існують також зв'язки еквівалентності між термінами на різних мовах. Виділяють повну (строгу) еквівалентність і декілька видів часткової (нестрогої) смислової еквівалентності термінів в різних мовах.

Тезаурус часто містить коментарі до термінів, що розкривають для користувача їх сенс, а також пояснюють, як слід використовувати його терміни.

Тезауруси застосовуються, перш за все, для класифікації і пошуку інформаційних ресурсів. При цьому кожному ресурсу при класифікації може бути співставлено одне або декілька понять, що описуються термінами в тезаурусі, а користувач, що здійснює пошук, може за тезаурусом знайти поняття, що цікавлять його в даній предметній області, а також всі терміни, що характеризують їх. Тобто на основі зв`язків тезауруса відбувається розширення пошукового запиту (розширення слів запиту синонімічними, більш загальними або більш приватними за сенсом термінами).

Існує низка стандартів різного рівня значущості і опрацьованості на формат представлення тезаурусів. Ці стандарти представляють тезаурус у вигляді набору об'єктів декількох типів, між якими може бути декілька типів зв'язків. Деякі стандарти (наприклад, стандарт ANSI/NISO Z39.19-1993) регламентують також формат представлення тезауруса в лінеаризованому (текстовому) вигляді, придатному для сприйняття як машиною, так і людиною.

Основними документами, що регламентують формат представлення тезауруса, є стандарти ISO 2788-1986 для опису одномовних тезаурусів і ISO 5964-1985 — для багатомовних.

Стандарт ISO 2788-1986 визначає тезаурус як набір термінів, пов`язаних між собою відповідними зв'язками (відносинами).

Структура багатомовних тезаурусів регламентується стандартом ISO 5964-1985. У ньому, крім зв'язків між термінами, є також зв`язки між еквівалентними термінами на різних мовах. Існують наступні типи таких зв'язків:

— повна еквівалентність;

— неповна еквівалентність (значення термінів не співпадають, але перетинаються);

— часткова еквівалентність (значення одного терміну ширше, ніж значення іншого);

— еквівалентність «один до багатьох» (значення одного терміну відповідає сукупності значень декількох термінів).

Американський стандарт ANSI/NISO Z39.19-1993 розширює і уточнює стандарт ISO 2788-1986 для одномовних тезаурусів, а також накладає додаткові обмеження на структуру тезауруса.

Інтерфейс перегляду тезауруса має:

— показувати всі атрибути даного терміна чи поняття;

— показувати, з якими термінами та поняттями пов'язано цей термін або поняття;

— досить наочно показувати користувачеві місце терміна чи поняття в ієрархії понять тезауруса.

Перші два пункти здійсненні, якщо показувати користувачу для кожного поняття тезауруса на окремому екрані (сторінці) всі його атрибути, всі пов'язані з ним терміни (на всіх або на певній мові), і всі пов'язані з ним поняття. Інтерфейс має при цьому забезпечувати перехід до сторінки перегляду будь-якого з перерахованих на цій сторінці понять. Якщо в тезаурус схемою даних дозволяється прив`язка терміну більш ніж до одного поняття, на тій же сторінці для кожного терміну мають бути перераховані також поняття, до яких ще прив'язаний цей термін. Якщо у поняття є терміни на інших мовах, не повністю еквівалентні даному поняттю або повністю еквівалентні, але прикріплені в силу структури даного тезауруса до інших понять, на сторінці мають бути присутні посилання на сторінки цих понять [11].

3. ПРОГРАМНА РЕАЛІЗАЦІЯ ПІДХОДУ

Для реалізації запропонованого підходу до автоматизованої побудови онтологій на основі знаходження відношень між термінами, зокрема відношення синонімії було розроблено програму, що складається з:

— модуля пошуку і виведення термінів;

— модуля аналізу термінів.

Припустимо, що тезаурус складається зі слів-визначень, які зберігається в окремому файлі, де назва файлу є слово-визначення (рис. 1), а вміст — всі його значення (рис. 2). Слова, які не відносяться до повнозначних частин мови, або є словами виключеннями: займенниками, числівниками, також зберігаються в окремому файлі. Словоформи слів теж зберігаються в окремому файлі.

Для нормальної роботи програми необхідний комп'ютер який задовольняє таким мінімальним системним вимогам:

— персональний комп'ютер на базі процесора Intel Pentium III і вище;

— ОС: Windows XP/Vista/Windows 7/ Windows 8;

— ОЗУ 256 Мб;

— вільне місце на жорсткому диску 100 Мб:

— наявність .NET Framework 3.5 SP 1.

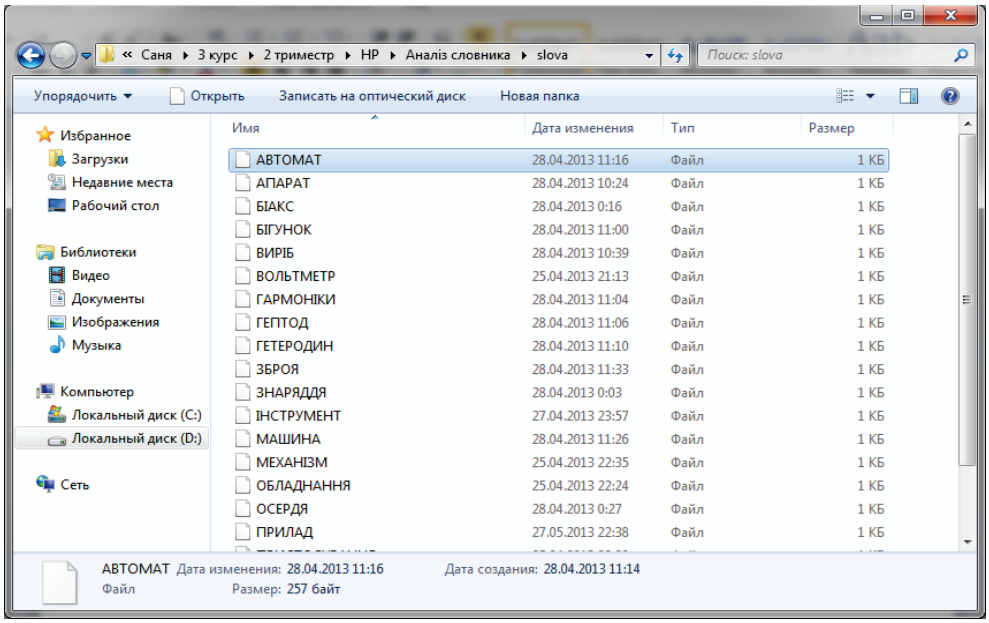


Рис. 1. Слова-визначення

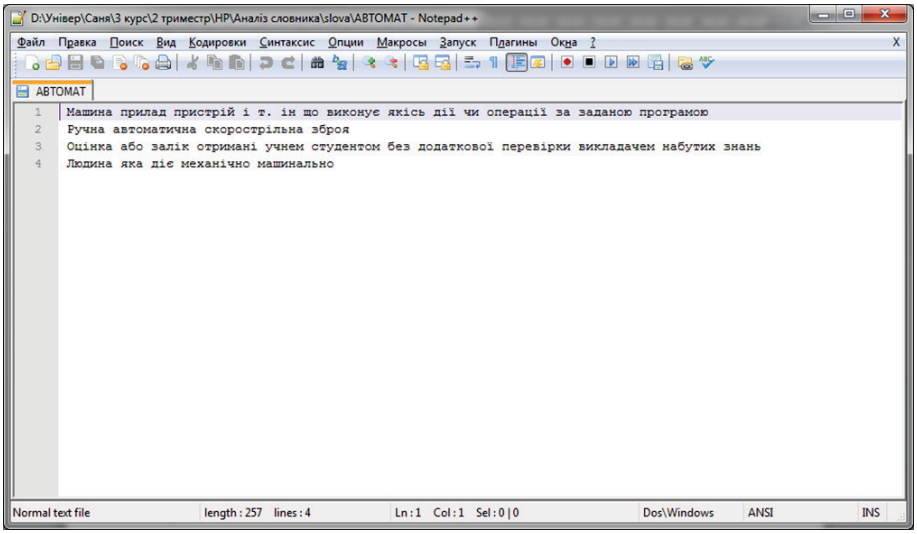


Рис. 2. Значення слова

Знаходження значення слова

Щоб знайти значення слова потрібно ввести його в строку пошуку і натиснути кнопку «Пошук». Після цього буде виведено всі значення слова або повідомлення, що таке слово не знайдене (рис. 3).

Пошук відбувається по директорії «slova», де зберігаються терміни за лінійний час, що пропорційний кількості файлів.

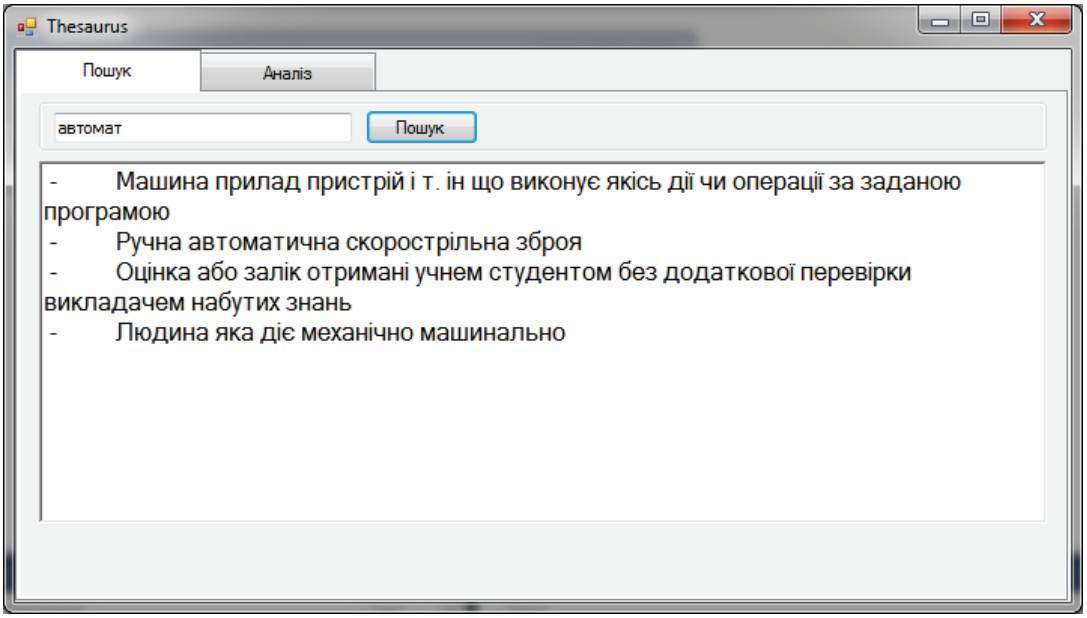
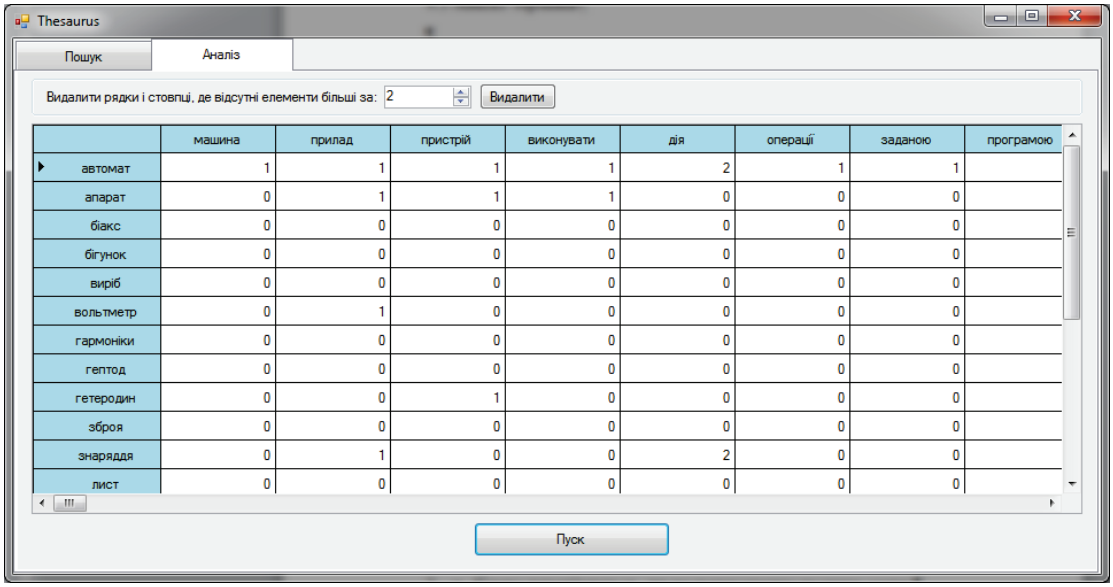


Рис. 3. Результати пошуку

Реалізація функції аналізу термінів

Аналіз термінів — це знаходження скільки разів те чи інше слово зустрічається у визначенні, що в свою чергу дає можливість знаходити контекстуальні синоніми до термінів.

Результатом роботи програми є таблиця, де рядки — це слова-визначення, а стовпці — слова, які містяться у визначенні. На перетині рядків і стовпців знаходиться цифра, що вказує скільки разів слово (у всіх його словоформах) зустрічається у даному визначенні (рис. 4). Також є можливість залишити рядки і стовпці утвореної таблиці, в яких є хоча б один елемент більший або рівний за число, вказане користувачем.



	машина	прилад	пристрій	виконувати	дія	операції	заданою	програмою
автомат	1	1	1	1	2	1	1	
апарат	0	1	1	1	0	0	0	
біакс	0	0	0	0	0	0	0	
бігунок	0	0	0	0	0	0	0	
виріб	0	0	0	0	0	0	0	
вольтметр	0	1	0	0	0	0	0	
гармоніки	0	0	0	0	0	0	0	
гептод	0	0	0	0	0	0	0	
гетеродин	0	0	1	0	0	0	0	
зброя	0	0	0	0	0	0	0	
знаряддя	0	1	0	0	2	0	0	
лист	0	0	0	0	0	0	0	

Рису. 4. Результат роботи модуля аналізу термінів

Побудова таблиці відбувається в декілька етапів:

а) кожне визначення розбивається на окремі слова;

б) слово перевіряється, чи це повнозначна частина мови;

в) якщо повнозначна — знаходиться його початкова форма, якщо не повнозначна — береться наступне слово і виконується пункт б;

г) перевіряється таблиця: якщо в заголовках стовпців це слово відсутнє — створюється стовпець з такою назвою і навпроти рядка з словом-визначенням ставиться 1, навпроти решти рядків — 0; якщо заголовок з таким словом є в таблиці — значення комірки на перетині відповідного рядка і стовпця інкрементується.

З утвореної таблиці, видаливши стовпці і рядки в яких відсутній хоча б один елемент більший або рівний за вказаний користувачем, отримаємо нову таблицю, в якій слово-значення зустрічається хоча б в одному визначенні більше або рівне раз заданого користувачем. З оновленої таблиці краще видно, які слова частіше зустрічається у визначенні слова (рис. 5).

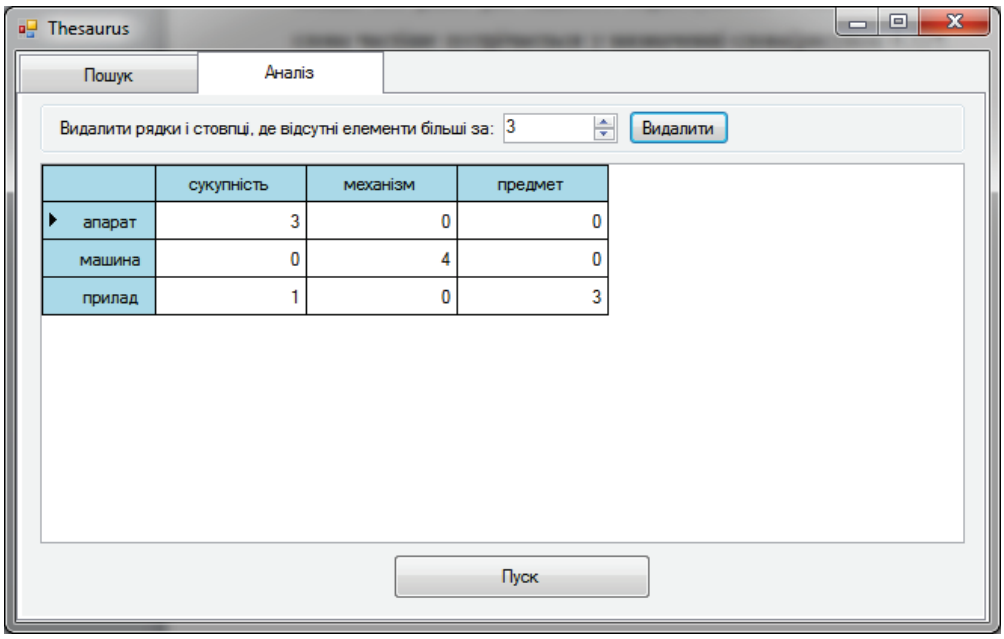


Рис. 5. Таблиця після видалення елементів

Розглянемо приклад. За допомогою ручного аналізу було отримано, що визначення вольтметр містить у собі слово прилад, яке містить — пристрій та інструмент. Пристрій, в свою чергу, складається з пристосування та обладнання. Пристосування містить пристрій і прилад (рис. 6).

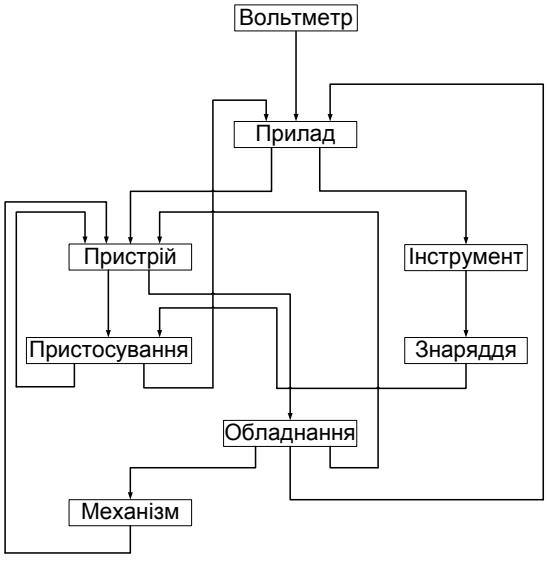
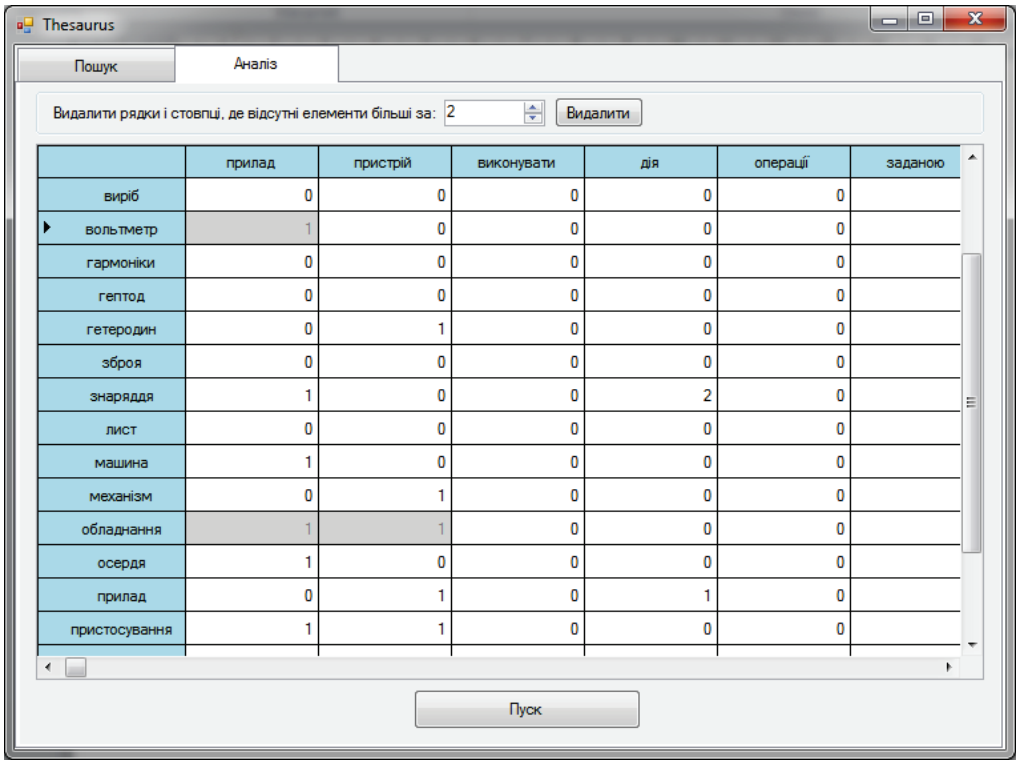


Рис. 6. Ланцюги термінів

За допомогою програмного аналізу отримано, що слово предмет у визначенні вольтметр зустрічається один раз (рис. 7). У визначенні прилад, пристрій зустрічається один раз, інструмент — два. Визначення пристрій містить пристосування та обладнання по одному разу. Пристосування містить пристрій і прилад по одному разу. Тобто результати програмного і ручного аналізу повність співпадають.



	прилад	пристрій	виконувати	дія	операції	заданою
виріб	0	0	0	0	0	
вольтметр	1	0	0	0	0	
гармоніки	0	0	0	0	0	
гептод	0	0	0	0	0	
гетеродин	0	1	0	0	0	
зброя	0	0	0	0	0	
знаряддя	1	0	0	2	0	
лист	0	0	0	0	0	
машина	1	0	0	0	0	
механізм	0	1	0	0	0	
обладнання	1	1	0	0	0	
осердя	1	0	0	0	0	
прилад	0	1	0	1	0	
пристосування	1	1	0	0	0	

Рис. 7. Результати програмного аналізу

ВИСНОВКИ

В статті розглянуто проблему автоматизованої побудови онтологій на основі знаходження відношень між термінами. Для досягнення мети дослідження запропоновано використовувати відношення синонімії за допомогою розробленої програмної процедури.

Також встановлено, що:

1. Модуль аналізу термінів дозволяє знаходити відношення синонімії у тексті та, на цій основі, автоматизовано будувати часткові онтології — у більшості існуючих систем такі відношення формуються вручну експертами-лінгвістами.

2. Специфікація та вимоги до побудови тезаурусу, на основі якого має працювати програмна процедура, відповідають традиційним лексикографічним продуктам, що збільшує предметну область застосування запропонованого підходу.

3. Результати отриманої програмної процедури у вигляді ланцюгів синонімів збігаються з аналогічними ланцюгами, що визначено вручну.

Отже, в даній роботі пропонується метод створення тезаурусу на основі вихідних даних та вбудованих лексикографічних знань тлумачного словника української мови. У розвиток запропонованого підходу та його програмній реалізації було б варто передбачити модуль графічного представлення отриманих онтологій на зразок рисунку 6.

СПИСОК ЛІТЕРАТУРИ

1. Широков В. А. Феноменологія лексикографічних систем / В. А. Широков. — К. : Наукова думка, 2004. — 327 с.
2. Широков В. А. Комп’ютерна лексикографія / В. А. Широков. — К. : Наукова думка, 2011. — 351 с.
3. Створення галузевого тезауруса в бібліотеці ХДУХТ. — [Електронний ресурс] — Режим доступу : http://blogs.mail.ru/mail/biblio-hduht/130B7-1BCA11854E2.html. — Назва з екрану.
4. Гулыга Е. В. Грамматико-лексические поля в современном немецком языке / Е. В. Гулыга.

Е. И. Шендельс. — М. : Просвещение, 1969. — 194 с.
5. Вердиева З. Н. Семантические поля в современном английском языке / Вердиева З. Н. — М. : Высшая школа, 1986. — 115 с.
6. Апресян Ю. Д. Лексическая семантика. Синонимические средства языка / Апресян Ю. Д. — М. : Наука, 1995. — 368 с.
7. Нижегородцева-Кириченко Л. А. Лексико-семантическое поле «ИНТЕЛЛЕКТУАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ» : опыт концептуального анализа (на материале существительных современного английского язика : дис. канд. филол. наук : 10.02.04 / Лариса Алексеевна Нижегородцева-Кириченко. — К., 2000. — 257 с.
8. Караулов Ю. Н. Общая и русская идеография / Караулов Ю. Н. — М. : Либроком, 2010. — 356 с.
9. Морковкин В. В. Идеографические словари / Морковкин В. В. — М. : Изд-во Моск. ун-та, 1970. — 71 с.
10. Кульчицький І. М. Розроблення WORDNET-подібного словника української мови / І. М. Кульчицький, А. Б. Романюк. // Вісник Національного університету «Львівська політехніка». Інформаційні системи та мережі. — 2010. — № 673. — С. 306—318.
11. Navarrete, E; Pastore, M; Valentini, R; Peressotti, P (2015). «First learned words are not forgotten: Age-of-acquisition effects in the tip-of-the-tongue experience». Memory & Cognition. 43 (7): 1085–1103. doi:10.3758/s13421-015-0525-3.
12. Рогушина Ю. В. Розробка онтологічних терміносистем інформаційних ресурсів інтернет та їх когнітивних моделей у наукових дослідженнях / Ю. В. Рогушина, А. Я. Гладун, В. Н. Штонда // Проблеми програмування. Спеціальний випуск — 2010. — № 2—3.
13. Тезауруси в описі інформації видавничої діяльності. — [Електронний ресурс]. — Режим доступу : http://www.big-library.com.ua/book/77_Informa-ciine_zabezpechennya_vidavnichoi_diyalnosti/7192_31_Tezayrysi_v_opisi_informacii_vidavnicho i_diyalnosti. – Назва з екрану.

SPUSOK LITERATURU

1. Shirokov V. A. Fenomenologíya leksikografíchnikh sistem / V. A. Shirokov. — K. : Naukova dumka, 2004. — 327 s.
2. Shirokov V. A. Komp’yuterna leksikografíya / V. A. Shirokov. — K. : Naukova dumka, 2011. — 351 s.
3. Stvorennya galuzevogo tezaurusa v bíblíotetsí KHDUKHT. — [Yelektronniy resurs] — Rezhim dostupu : http://blogs.mail.ru/mail/biblio-hduht/130B7-1BCA11854E2.html. — Nazva z yekranu.
4. Gulyga Ye. V. Grammatiko-leksicheskiye polya v sovremennom nemetskom yazyke / Ye.V. Gulyga. Ye.I. Shendel's. — M. : Prosveshcheniye, 1969. — 194 s.
5. Verdiyeva Z. N. Semanticheskiye polya v sovremennom angliyskom yazyke / Verdiyeva Z. N. — M. : Vysshaya shkola, 1986. — 115 s.
6. Apresyan YU. D. Leksicheskaya semantika. Sinonimicheskiye sredstva yazyka / Apresyan YU. D. — M. : Nauka, 1995. — 368 s.
7. Nizhegorodtseva-Kirichenko L. A. Leksiko-semanticheskoye pole «INTELLEKTUAL'NAYA DEYATEL'NOST'»: opyt kontseptual'nogo analiza (na materiale sushchestvitel'nykh sovremennogo angliyskogo yazika: dis. kand. filol. nauk : 10.02.04 / Larisa Alekseyevna Nizhegorodtseva-Kirichenko. — K., 2000. — 257 s.
8. Karaulov YU. N. Obshchaya i russkaya ideografiya / Karaulov YU. N. —M. : Librokom, 2010. — 356 s.
9. Morkovkin V. V. Ideograficheskiye slovari / Morkovkin V. V. — M. : Izd-vo Mosk. un-ta, 1970. — 71 s.
10. Kul'chits'kiy Í. M. Rozroblennya WORDNET-podíbnogo slovnika ukraíns'koí movi / Í. M. Kul'chits'kiy, A. B. Romanyuk. // Vísnik Natsíonal'nogo unívesitetu "L'vívs'ka polítekhníka". Ínformatsíyní sistemi ta merezhí. — 2010. — № 673. — C. 306—318.
11. Navarrete, E; Pastore, M; Valentini, R; Peressotti, P (2015). «First learned words are not forgotten: Age-of-acquisition effects in the tip-of-the-tongue experience». Memory & Cognition. 43 (7): 1085–1103. doi:10.3758/s13421-015-0525-3.
12. Rogushina YU. V. Rozrobka ontologíchnikh termínosistem ínformatsíynikh resursív ínternet ta íkh kognítivnikh modeley u naukovikh doslídzhennyakh / YU. V. Rogushina, A. YA. Gladun,

V. N. Shtonda // Problemi programuvannya. Spetsíal'niy vipusk — 2010. — № 2—3.
13. Tezaurusi v opisí ínformatsíí vidavnichoí díyal'ností. — [Yelektronniy resurs]. — Rezhim dostupu : http://www.big-library.com.ua/book/77_Informa-ciine_zabezpechennya_vidavnichoi_diyalnosti/7192_31_Tezayrysi_v_opisi_informacii_vidavnichoi_diyalnosti. – Nazva z yekranu.

Надійшла до редакції 2.12.2016 р.

БІСІКАЛО О. В. — д.т.н., проф. каф. АІВТ, декан ФКСА, Вінницький національний технічний університет, м. Вінниця, Україна.

ЯХИМОВИЧ О. В. — аспірант каф. АІВТ, Вінницький національний технічний університет, м. Вінниця, Україна.