УДК 81.374

І.М. Кульчицький, А.Б. Романюк, Х.Б. Харів
Національний університет “Львівська політехніка”,
кафедра систем автоматизованого проектування

РОЗРОБЛЕННЯ WORDNET-ПОДІБНОГО СЛОВНИКА УКРАЇНСЬКОЇ МОВИ

Проаналізовано організацію іменників у словнику WordNet, виділення сукупності іменників, що становлять основу лексичної бази даних української мови, упорядкування цих іменників на основі лексичних і семантичних зв’язків з утворенням ієрархічних структур, реалізація наявних зв’язків та побудова WordNet-подібного словника для частини іменників сформованої сукупності.

Ключові слова: словник, WordNet, синсет, зв’язки гіперонімії, гіпонімії, антонімії; іменник, основні концепти, ієрархія.

The research analyzes arrangement of nouns in the Princeton WordNet, namely, the top concepts and key hierarchies;. establishes a selection of nouns constituting the core lexicon of the Ukrainian language; organizes the selected nouns into a hierarchical structure by means of semantic and lexical relations; implements the relations between nouns to build a WordNet-like dictionary containing core concepts.

Keywords: dictionary, WordNet, synset, relations hypernymy, hyponymy, antonymy, noun, base concepts, hierarchy.

Вступ

Опрацювання природної мови сьогодні належить до найпрогресивніших і найактуальніших завдань комп’ютерної лінгвістики. Правильне розуміння мови можливе за умови наявності знань про те, як слова та поняття пов’язані між собою, що мається на увазі під тим чи іншим висловлюванням, що мовець має на меті, кажучи ту чи іншу фразу; що сказано, а що необхідно віднайти в контексті або сприйняти на основі попередньо засвоєної інформації. Така комплексність завдання створює значні перешкоди для автоматизації роботи з природними мовами, однак, зважаючи на важливість подолання цього бар’єра, дослідження, спрямовані на створення систем, що охоплюють принаймні частину аспектів сприйняття, аналізу та розуміння мови, невпинно продовжуються, і поступ у цій галузі очевидний.

На вирішення завдань аналізу відношень між словами та поняттями, концептуалізації дійсності тощо спрямовано розроблення так званих лексичних та лексико-семантичних баз даних або структурованих лексиконів. До таких систем належать Прінстонський WordNet, MindNet, програмний продукт проекту дослідницького відділу Майкрософт, FrameNet, розроблений в університеті Берклі, VerbNet, HowNet, ConceptNet тощо.

ConceptNet – семантична база даних, створена в Массачусетському технічному інституті, в якій зібрано так звані загальні знання, що ними володіють люди і несвідомо послуговуються для сприйняття мови. У цій системі реалізовано 20 типів семантичних відношень на зразок «причина», «розташування», «використовують для» тощо, а найуживаніші концепти організовані у вигляді графів [1].

HowNet – лінгвістичний ресурс, що складається з дев’яти баз даних, у яких поєднано знання, що їх носії мови використовують для щоденного спілкування, та лексична інформація. Поняття у системі HowNet пов’язані між собою 20 типами семантичних відношень [2].

MindNet від дослідницького відділу компанії Майкрософт – інший приклад лексико-семантичної бази даних великого обсягу, побудованої у вигляді системі графів, у якій реалізовано 15 семантичних відношень [3].

Інший підхід було використано під час розроблення системи FrameNet, до складу якої входить лексикон, база даних фреймів, кожен із яких є схематичним відтворенням певної ситуації, та речення-приклади [4].

Лексико-семантичну базу даних дієслів англійської мови VerbNet розробили фахівці університету Колорадо. Це ієрархічно організований загальний дієслівний лексикон великого обсягу, побудований на основі класів дієслів за Левін [5], що, зокрема, підтримує рольові та тематичні відношення [6].

Прінстонський WordNet, перша версія якого була створена ще в червні 1991 року, – це лексико-семантична база даних англійської мови, що характеризується великим обсягом, точністю та надійністю даних [7].

1. Постановка проблеми

Усі згадані лінгвістичні ресурси допомагають вирішити завдання встановлення значення омонімів, полісемантичних слів, автоматизованого підсумування текстів, розширення можливостей пошукових систем, визначення настрою, погляду чи намірів автора документа тощо. Цілком зрозуміло, що кожен із них має як переваги для розв'язання одних прикладних задач, так і обмеження в інших ситуаціях. Однак особливо широкого застосування набув саме WordNet, який сьогодні вважають одним із найнеобхідніших і найуніверсальніших інструментів у галузі комп'ютерної лінгвістики.

Розроблення WordNet-подібного словника для української мови сприятиме розвиткові галузі опрацювання природної мови та лінгвістики загалом, адже тезаурус такого типу є також надійним джерелом інформації про конкретну мову.

2. Аналіз останніх досліджень та публікацій

Робота над створенням структурованого лексикону, який поєднував би переваги традиційних комп'ютерних словників та тезаурусів, почалася в Прінстонському університеті ще в 1985 році. Керуючись найновішими тогочасними здобутками психолінгвістики, команда дослідників на чолі з професором А. Дж. Міллером поставила собі за мету створити базу даних англійської мови. WordNet і досі залишається найбільшим за обсягом програмно реалізованим лексиконом. Попри початкову психолінгвістичну спрямованість досліджень, широке застосування WordNet знайшли саме фахівці комп'ютерної лінгвістики. WordNet – потужний інструмент, який використовують для вирішення різноманітних завдань галузі опрацювання природної мови, адже він надає можливості розробки дієвих алгоритмів встановлення значень полісемантичних слів, що є ключем до вирішення цілої низки прикладних лінгвістичних завдань. WordNet застосовують для автоматизованого реферування та категоризації текстів, добування інформації, генерування лексиконів, створення засобів автоматизованого перекладу та пошукових систем тощо.

Специфікою WordNet є спроба його розробників організувати лексичну інформацію, беручи за основу насамперед значення слова, а не його форму. У цьому аспекті WordNet ближчий до тезауруса, аніж до словника. Однак, на відміну від традиційних тезаурусів, WordNet (хоча він і характеризується певною надлишковістю входжень, зумовленою потребою щоразу експлікувати ту чи іншу лінгвістичну інформацію) влаштовано у такий спосіб, що зазначені процедури, завдяки можливостям комп'ютерів, виконуються миттєво, а результат пошуку є наочним і повним. Так, якщо в друкованому виданні алфавітного тезауруса X і Y є синонімами, то ця пара повинна бути наведеною двічі – на X і на Y, а в тематичному тезаурусі її пошук взагалі вимагатиме двох процедур – спочатку в алфавітному списку, а потім власне у тезаурусі – цих кроків користувач WordNet взагалі не помічає.

Проте WordNet – це не простий діалоговий тезаурус. Щоб оцінити його переваги, треба виходити з того, що лексична семантика починається із розуміння факту, що слово реалізує загальноприйняту асоціацію між лексикалізованим поняттям та синтаксично зумовленим висловлюванням. Це визначення зумовлює появу комплексу з трьох аспектів дослідження: по-перше, які види висловлювань вступають у ці лексичні асоціації; по-друге, що являють собою і як організовані поняття, виражені словами; і по-третє, яку синтаксичну роль виконують різні слова. Не ігноруючи жодної з цих трьох проблем, теорія WordNet зосереджується на другому колі питань, що має безпосереднє відношення до семантичної структури англійського лексикону (як і лексикону будь-якої іншої мови).

Деякі труднощі викликає термінологія, якою послуговуються розробники та дослідники WordNet, зокрема, використання англійських слів ‘concept’, ‘word’, ‘word form’ тощо. Оскільки "слово", зазвичай, вживають для позначення як висловлювання, так і його асоціативного поняття, виникає певна термінологічна двозначність. Задля її уникнення під формою слова (чи словоформою) надалі слід розуміти позначення «фізичного» висловлювання чи напису, а під значенням слова – позначення лексикалізованого поняття, для виразу якого використовується ця форма. Такий спрощений підхід до дослідження лексичної семантики полягає у встановленні відповідності між формами та значеннями слів.

Якщо зафіксовано принаймні два різні тлумачення (у «правій» частині словникової статті) для однієї словоформи, то вона є полісемантичною (за винятком випадків омонімії); якщо ж для різних словоформ подають однакові інтерпретації, то вони є синонімічними. З погляду психолінгвістики полісемія і синонімія виявляють проблеми, зумовлені пошуком відповідного доступу до інформації в ментальному лексиконі [9].

У своїй «класичній» концепції WordNet має справу винятково зі встановленням семантичних відношень між лексикалізованими поняттями, тобто WordNet можна кваліфікувати як теорію поля значення слова. Проте не можна ігнорувати також лексичні відношення між формами слова, адже вони роблять свій внесок в організацію WordNet. Оскільки семантичні відношення – це відношення між значеннями та оскільки синонімічні ряди репрезентують певне значення, природно вважати, що зв’язки між синонімічними рядами відображають семантичні відношення. Характерно, що семантичні відношення симетричні, тобто, якщо існують відношення R між значенням {x, x', ...} і значенням {y, y', …}, то є також відношення R' між {y, y', …} і {x, x', ...}. У межах обговорюваної теми імена семантичних відношень відіграватимуть подвійну роль: якщо R – відношення між значеннями {x, x`...} та {y, y`, ...}, тоді R також застосовуватиметься для позначення відношень між окремими формами слова, які належать до тих самих синонімічних рядів [9].

Саме синонімічні ряди є основною складовою одиницею словника WordNet. У міжнародній термінології їх прийнято називати синсетами. Усі синсети пов’язані між собою системою семантичних зв’язків; у словниках WordNet реалізовані також лексичні зв’язки, що сполучають окремі слова – члени синонімічних рядів. За початковою ідеєю, лексичну базу даних мали формувати лише набори синонімів – де кожен відображає певний концепт – що утворюють систему зв’язків. Однак, як виявилося, самих синсетів, що фактично є рядом контекстуальних синонімів, та лексичних і семантичних відношень, які відіграють роль контексту, зазвичай недостатньо для повного та вичерпного визначення певного поняття. Саме тому розробники WordNet вирішили ввести до складу синсетів короткі тлумачення (glosses), а також зразки вживання.

Лексикон, розроблений дослідниками Прінстонського університету, охоплює чотири частини мови: іменники, дієслова, прикметники та прислівники. Для кожної із цих синтаксичних категорій створено окремі файли, що істотно відрізняються структурною організацією. Причому спосіб структурування кожної з синтаксичних категорій є індивідуальним. Так, іменники організовано у лексичній пам’яті як тематичні ієрархії, дієслова систематизовано на основі імплікативних відношень, а прикметники і прислівники – як N-вимірні гіперпростори із використанням деривaційних відношень. Кожна із зазначених структур відображає різний шлях і спосіб категоризації мовного матеріалу. Спроби накладання єдиного принципу організації на всі синтаксичні категорії не є релевантними, оскільки вони спотворюють психологічну складність лексичного знання. Тому підсистеми мовної системи будуються в межах певної частини мови, що зумовлює специфіку організації лексикографічних структур. Частиномовну стратифікацію мовної системи підтверджує й експериментальне вивчення словесних асоціацій [9].

WordNet побудований завдяки явищу синонімії, яка є засадничим відношенням цього семантичного словника. Однак варто зазначити, що не йдеться про абсолютну синонімію, оскільки слова, що є абсолютними синонімами, практично не існують в природних мовах, а на думку деяких дослідників, узагалі не існують. У WordNet зафіксовано випадки контекстуальної синонімії, тобто

два слова є синонімічними в контексті С, якщо заміна одного на інше в С не змінює його істинності. Отже, слова, що утворюють синсет, є взаємозамінними лише в певному контексті.

Отже, синонімічні ряди можуть утворювати лише слова, що належать до однієї частини мови, а слова різних синтаксичних категорій не можуть бути синонімами. Відповідно використання синонімічних рядів для представлення значень слів відображає психолінгвістичні положення, які свідчать про те, що іменники, дієслова і модифікатори організовані в семантичній пам'яті незалежно. Підтвердженням цього є той факт, що деякі слова однієї і тієї самої синтаксичної категорії (переважно дієслова) виражають дуже близькі поняття, однак вони не можуть бути взаємозамінними, оскільки це зумовлює виникнення граматичних помилок.

Відношення синонімії дає змогу сформувати основні складові одиниці словника WordNet – синсети (синонімічні ряди). Однак для того, щоб організувати їх у цілісну систему, використано низку інших семантичних (гіпо-/гіперонімія, меронімія та голонімія, тропонімія тощо) та лексичних відношень (антонімія).

У систематизації іменників найважливішу роль відіграє родо-видове відношення, відоме як гіперонімія, та обернене йому відношення гіпонімії. Саме завдяки цим семантичним відношенням іменники можна організувати з утворенням ієрархічних структур. Кожне загальне поняття розбивається на конкретніші значення, що утворюють розгалуження. Своєю чергою, кожен гіпонім, представлений конкретною словоформою, має один безпосередній гіперонім, що дає змогу встановити значення цієї словоформи, відрізняючи його від інших можливих варіантів. Це дає змогу подолати омонімію та багатозначність слів. Так, для прикладу, значення слова «стан» залежить від його гіпероніма: якщо йдеться про сукупність ознак, то загальнішим поняттям буде «характеристика»; якщо ж мається на увазі корпус людини, безпосереднім гіперонімом буде «частина тіла». Однією із характерних особливостей WordNet є виділення найтонших відтінків значення, кожен із яких входить до складу окремої ієрархії.

Важливою властивістю семантичних ієрархій, побудованих на основі гіпо-/ гіперонімії, є спадковість (спадкування). Гіпонім вбирає в себе усі риси загального поняття і додає риси, що відрізняють його від гіпероніма і від інших гіпонімів цього гіпероніма. Це дає змогу забезпечити центральний принцип організації іменників у системі WordNet, а також уникнути дублювання інформації.

Іншим типом відношень є "частина-ціле", або так звана меронімія. Поняття, представлене набором {x, x', ...}, є меронімом поняття, представленого набором {y, y', ...}, якщо носій мови сприймає речення, сконструйоване з такої основи, як: Y має X (як частину) або ж X є частиною Y . Це відношення також інверсійне, оберненою до нього є голонімія. Меронімія та голонімія доповнюють ієрархії, сформовані на основі родо-видових відношень, внаслідок чого виникають складні системи. Це семантичне відношення також підпорядковується принципу спадкування. Особливо важливу роль меронімія/голонімія відіграє для організації іменників, що позначають частини тіла та інших природних об'єктів, деталі предметів, створених людиною, а також величини.

Необхідно зауважити, що зв'язок між поняттями, пов'язаними відношенням меронімії/голонімії, не завжди можна описати конструкцією «має частину/є частиною». У WordNet виділяють три типи меронімії, які прийнято розуміти так: власне «є частиною», «є представником» та «є речовиною, з якої складається». Так, наприклад, двері є частиною будинку, батько є представником сім'ї, а тканина є речовиною, з якої складається частина тіла. Як стверджують дослідники, відношення «є частиною» є найпоширенішим різновидом меронімії.

На відміну від гіпонімії та меронімії, антонімія пов'язує не поняття, а конкретні словоформи. Антонімія не забезпечує утворення ієрархічних структур. Однак важливість цього лексичного відношення полягає в тому, що – як показали психолінгвістичні дослідження – слова, що мають протилежне значення, пов'язані в пам'яті людини. Так, якщо провести тест на асоціації, то найімовірніше, що спонтанною відповіддю на подане слово буде саме його антонім [8]. Отже, відображення відношення лексичної опозиції забезпечує виконання вихідних теоретичних засад WordNet.

Варто пам'ятати, що антонімія пов'язує не поняття, а лише певні словоформи, тобто не цілі синонімічні ряди, а конкретні їх члени. Крім того, це відношення не має властивості спадкування, тому антонімію між словоформами, що позначають відповідні гіпо- та гіперонімічні поняття,

потрібно заносити окремо. Ще одна цікава особливість антонімічних словоформ полягає в тому, що вони, зазвичай, мають спільні безпосередні гіпероніми [7].

Одним із випадків антонімії є так звана градація, коли антонімічний набір формують словоформи, що позначають не лише протилежні, але й проміжні поняття. Цей тип відношення притаманний здебільшого прикметникам (напр., холодний, теплий, гарячий), однак він зустрічається і серед іменників (дитинство, юність, старість; початок, середина, кінець).

Як уже було вказано раніше, кожна синтаксична категорія у WordNet організована по-іншому. Відповідно, тоді як іменники систематизовано завдяки гіпо-/гіперонімії, меронімії/ голонімії та антонімії, дієслова, прикметники та прислівники пов'язані між собою іншими типами лексичних та семантичних відношень.

4. Формулювання цілей статті

У ході реалізації міжнародних проектів на зразок EuroWordNet [10] та BalkaNet [11], а також численних локальних досліджень було розроблено низку методів побудови комп'ютерних тезаурусів, структура яких відповідає принципам побудови WordNet. Існує два основні підходи до створення тезаурусів такого типу, що відрізняються підходом до відображення семантичної системи досліджуваної мови.

Метод входження (під'єднання, приєднання) (the merge model) полягає у побудові лексикону певної мови на основі попереднього ґрунтовного аналізу її семантичної системи і подальшому встановленні відповідностей із англійським WordNet та так званим міжмовним індексом (Interlingual index, ILI) на основі зв'язку еквівалентності.

Метод розширення (the expand model), навпаки, передбачає переклад ієрархій, що їх виділяють в англійському чи іншому наявному семантичному словникові, досліджуваною мовою та відтворення відповідних зв'язків.

Обидва підходи мають як сильні, так і слабкі сторони. Зокрема, результатом реалізації методу розширення може стати створення копії словника – прототипу, що не відображає реальної структури мови. Метод входження, з іншого боку, дає змогу змоделювати семантичну систему, враховуючи особливості лексикалізації та концептуалізації досліджуваної мови, проте реалізувати його значно складніше, з огляду на комплексність та масштабність завдання такого характеру. Крім цього, застосування перекладного методу гарантує відповідність будови отриманого словника структурі прототипу, тоді як лексикон, побудований без урахування особливостей семантичної організації наявних WordNet-подібних словників, може містити істотні відмінності, що стане перешкодою для встановлення відповідностей із системою словників wordnet та загальноприйнятими онтологіями.

Існує також кілька підходів до виконання одного із початкових завдань, необхідних для створення лексикону за взірцем WordNet, – формування сукупності так званих базових концептів (Base Concepts), тобто понять, що становлять основу лексики досліджуваної мови. Основні концепти визначають здебільшого за допомогою частотних словників або корпусів текстів, що охоплюють усі мовні стилі.

Останнім часом популярності набуває інший підхід – асоціативний. Він полягає в опитуванні носіїв мови, яким пропонують певний перелік слів, до яких потрібно подати асоціації. Метод вільних асоціацій (FAT, 'free association test'), як його прийнято називати, дає надійні результати, однак його втілення передбачає виконання попереднього психолінгвістичного дослідження.

Методи розроблення лексиконів досліджуваної моделі відрізняються також використанням засобів побудови. WordNet-подібні словники можуть бути створені вручну або ж із використанням комп'ютерних технологій для автоматизації того чи іншого етапу розроблення. Зрозуміло, що використання програмних засобів дає змогу пришвидшити процес побудови, однак такий підхід не гарантує точності. Під час автоматизованої розробки помилок уникнути не вдасться, тому цей підхід передбачає здійснення подальшої ручної перевірки та внесення відповідних виправлень. Натомість ручна побудова WordNet-подібного словника, за умови, що над нею працює команда кваліфікованих лінгвістів, забезпечує належне виконання усіх завдань. Загалом, оптимальним є використання комп'ютерних технологій з подальшим ручним обробленням отриманих результатів.

Варто зауважити, що вибір методів побудови словника wordnet переважно здійснюється на основі лінгвістичних ресурсів, доступних для досліджуваної мови, а також кадрову та фінансову забезпеченість. Так, для автоматизації процесу розроблення – залежно від завдань, які виконуються з використанням комп'ютерних технологій, – необхідні розмічені одно-, двомовні корпуси текстів, морфологічні аналізатори, а також фахівці галузі інформаційних технологій, здатні створити відповідні програмні засоби. Застосування згаданого вище методу входження, своєю чергою, неможливе без наявності лексикону досліджуваної мови. Отже, переваги та недоліки відходять на задній план, а методи розроблення вибирають з огляду на доступні ресурси та засоби.

Зважаючи на відсутність не лише лексикону, а й ґрунтовних досліджень семантичних та лексичних зв'язків між іменниками в українській мові, метод входження не може бути застосованим для розроблення українського WordNet-подібного словника. Отже, будувати семантичний словник для української мови варто, беручи за основу модель розширення. Тут слід зауважити, що деякі спроби втілити цей підхід уже зроблено на кафедрі систем автоматизованого проектування Національного університету «Львівська політехніка». Бакалаврам прикладної лінгвістики дали завдання перекласти іменникові ієрархії, побудовані завдяки семантичним зв'язкам гіпо- та гіперонімії. Спробу навряд чи можна вважати вдалою, хоча такий експеримент показав усі недоліки методу розширення. Насамперед, необхідно підкреслити, що цей підхід передбачає не переклад, а визначення українських еквівалентів відповідних англійських понять. Із цим завданням студенти, вочевидь, не впоралися, оскільки не ознайомилися з фундаментальними принципами словника WordNet. Така спроба підтвердила, що модель розширення можна застосовувати лише частково, враховуючи особливості досліджуваної мови, сприйняття тих чи інших понять носіями мови, способи лексикалізації, лексичні, граматичні та культурні реалії тощо. Реалізація методу розширення повинна здійснюватися лінгвістами, обізнаними в питаннях способу організації тезаурусних WordNet-подібних словників, результати мають підлягати ретельній перевірці, для того щоб забезпечити уникнення помилок, неточностей та суб'єктивності. Отже, можна зробити висновок, що описані вище підходи доповнюють один одного, їх слід поєднувати, що дасть змогу використати переваги обох методів й оминути недоліки. Крім того, варто починати створення національного wordnet зі структурної організації обмеженої кількості лексичних одиниць, а саме найуживанішої лексики, тобто слів, що формують понятійну основу мови.

5. Виклад основного матеріалу

Розроблення словника Wordnet для певної мови – проект, яким займається команда щонайменше п'яти лінгвістів, зокрема лексикографів, упродовж трьох років за умов відповідного фінансування. Тому, з огляду на обмежений час та засоби, було вирішено звузити завдання дослідження до розробки WordNet-подібного словника для іменників української мови.

Початковий етап побудови українського wordnet – вибір основних концептів, а саме відповідних іменників. Автоматизовано здійснити цю операцію було неможливо, оскільки отримати корпус текстів для української мови не вдалося. Відповідно, слова, що становлять основу української мови, вибирали за принципом частотності. З метою отримання достовірної інформації було використано частотні словники лексики різних літературних стилів (художня проза, публіцистика) та різних років видання (від 1980 до 2005 р.). Спочатку було сформовано чотири окремі списки найчастотніших іменників (для отримання переліку близько 1000 одиниць встановлена мінімальна абсолютна частота – 18). Необхідно зауважити, що чимало іменників було вилучено через стилістичні особливості (дівка, заграва, уста, дідько), застарілість (воєвода, шляхтич, десятина, дукач) тощо; до списку не ввійшли також відверті русизми (гімнастьорка, чемодан, грузовик, жених тощо), слова, притаманні тому часовому періоду, що його відображають твори, використані під час складання частотних індексів того чи іншого словника (комсомолець, комсомол, райком, фронт, татарин, більшовик, колгосп, поранені, полонені; маршрутка, криза, недовіра, нафта). Інші лексичні одиниці відображають жанрові особливості літератури (плісень, цебер, бомба у списках художньої прози; пленум, комбанк, пільга, легітимність, прес-конференція, прес-служба, опозиція, фальсифікація, приватизація у публіцистиці, особливо найновішій).

Щоправда, вилучені іменники усе ж було враховано при визначенні найчастотніших загальних понять (соціальна група, професія, посада, частина тіла, одяг тощо).

Лексичні одиниці, які потрапили щонайменше у два переліки, було відібрано до загального списку. Це забезпечило відбір іменників, які становлять основу лексикону, незалежно від жанрових особливостей у різні історичні періоди.

Використання частотних словників необхідне для виконання однієї з трьох основних вимог до лексикону – відображення лексичних одиниць, притаманних досліджуваній мові.

Зважаючи на ієрархічну організацію іменників у словниках wordnet, до підсумкової сукупності було також введено загальні поняття, отримані за допомогою поверхневої класифікації найчастотніших лексичних одиниць (емоція, почуття, людина, професія, сім'я, будівля, транспортний засіб, одяг тощо). Таке групування має на меті забезпечити іншу важливу властивість лексикону – узагальненість.

Однак заздалегідь передбачити всі загальні поняття, необхідні для утворення ієрархічних структур, надзвичайно складно. До того ж, WordNet-подібні словники містять чимало технічної лексики, новотворів та штучно введених понять, впровадження яких, безумовно, не відповідає принципові відображення організації ментального лексикону. Наявність таких лексичних одиниць, однак, необхідна для забезпечення ієрархічності і знову ж таки забезпечує узагальненість та всеохопність лексикону. Визначити необхідні загальні поняття, а також штучні та технічні лексичні одиниці, які необхідно додати до основних концептів, дає змогу аналіз сукупності базових концептів, сформованої під час розробки англійського WordNet 1.5 (так звані Base Concepts 1, загальноприйняте скорочення – BC1). Необхідно зазначити, що наступні версії прінстонського тезауруса характеризуються істотними структурними змінами, проте набір базових концептів для оновлених словників не подають. Тому потрібно прослідкувати зміни виділення ієрархічних вузлів у найновішій версії.

Під час визначення лексичних одиниць, які мають входити до складу базових понять, враховують не лише частоту вживання, а й кількість зв'язків, що поєднує певне слово з іншими. Значна частина слів, для яких виконується друга умова, утворюють верхні рівні ієрархії. Для прикладу, сьогодні найвищі рівні ієрархії іменників WordNet мають такий вигляд (рис. 1).

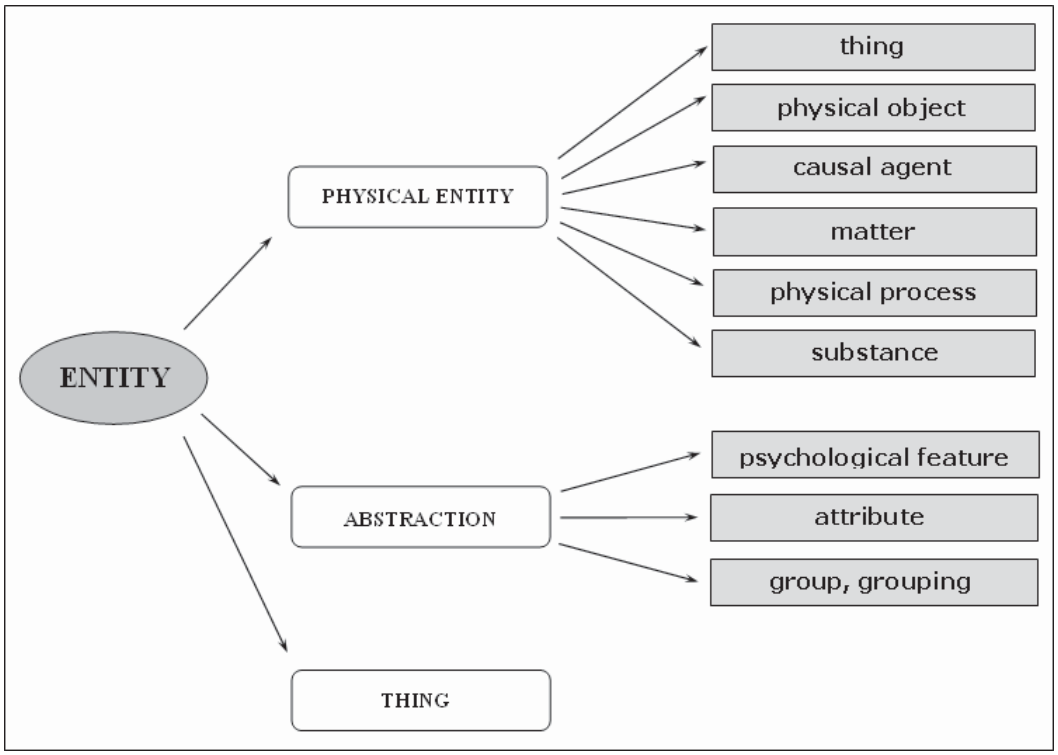

Рис. 1. Найвищий рівень іменникової ієрархії WordNet 3.0

До найпоширенішої технічної лексики, необхідної для побудови семантичного словника за взірцем WordNet, належать назви таксонів – одиниць біологічної класифікації живих істот (евкаріоти, хребетні, хордові, плацентарні тощо), фізичні поняття (матерія, речовина, субстанція тощо) тощо.

Вибір базових концептів – лише перший крок на шляху до створення словника wordnet. На наступному етапі сформовану сукупність необхідно впорядкувати з утворенням системи, що об'єднувала б усі іменники. Оскільки категоризація найвищих рівнів має порівняно універсальний характер, структура ієрархії найвищого порядку частково відповідає способу організації WordNet 3.0 (див. рис. 2).

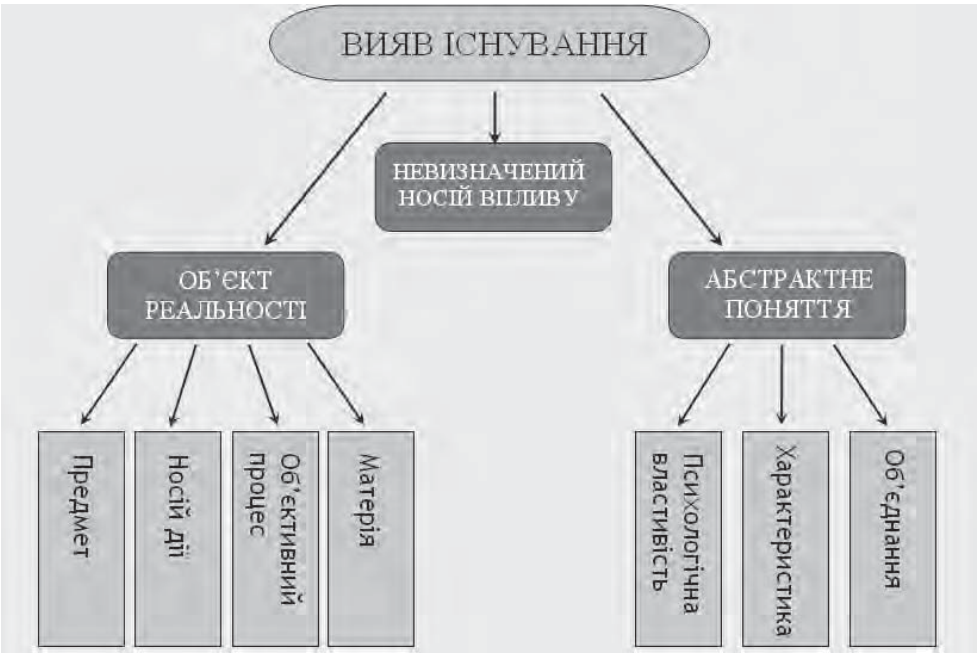

Рис. 2. Найвищий рівень іменникової ієрархії WordNet-подібного словника української мови

Оскільки еквівалента англійському "entity", що охоплює абсолютно всі іменники і є єдиним похідним поняттям, знайти не вдалося, було використано новотвір "вияв існування". Як видно з рис. 2, усі поняття поділено на дві групи відповідно до того, чи вони позначають об'єкти дійсності, що існують поза межами людської уяви і не залежать від неї, чи такі, що є продуктом діяльності людського мозку (об'єкт реальності та абстрактне поняття, відповідно). Третя група – {невизначений носій впливу, невизначений носій властивості} – є штучно введеним еквівалентом англійського 'thing', що об'єднує узагальнені поняття, що їх використовують для позначення як реальних, так і уявних об'єктів, які мають специфічні властивості або особливий вплив (головно, на людей). Українською мовою таке значення передаємо здебільшого неозначеним займенником 'щось'. Зважаючи на те, що цей вузол ієрархії не відіграє важливої ролі в загальній організації іменників, питання визначення найвідповіднішого еквівалента на цьому етапі не є принципово важливим.

Отже, найвищий рівень утвореної ієрархії українських іменників майже повністю повторює структуру прінстонського прототипу. Однак організація наступних вузлів уже потребувала деяких змін. Так, зокрема, у Прінстонському WordNet "physical entity" має шість гіпонімів: "thing", "object, physical object", "causal agent, cause agency", "matter", "process, physical process", і "substance" (річ, предмет, носій дії, матерія, процес і субстанція). Проте в українській мові слово «річ» у своєму найзагальнішому значенні та «предмет» є синонімами. Схожа ситуація з поняттями «субстанція», яке словники тлумачать як "будь-яка речовина взагалі", та «матерія», що має таке значення: «те, з чого складаються всі тіла в природі». Відповідно, як зображено на рис. 2, українське поняття "об'єкт, об'єкт реальності", має чотири гіпоніми: "предмет", "носій дії", "об'єктивний процес" і "матерія".

Поняття, що формують найвищий рівень ієрархії, доволі часто не лексикалізовані, тому їх організація виглядає дещо штучно, проте це необхідна умова створення єдиної системи. Натомість мета подальшої систематизації – побудувати «каркас», саму структуру, яку згодом можна було б заповнити елементами, тобто синсетами, що позначають відповідні поняття, насамперед найужи-

ваніші, що, власне, і належать до сукупності базових понять. Отже, завдання зводиться до формування найважливіших гіпо- / гіперонімічних послідовностей, що дають змогу розмістити базові поняття в загальній ієрархії.

Необхідно пам'ятати, що найуживаніші слова зазвичай полісемантичні, а тому в загальну систему потрібно вписати всі найважливіші їхні значення. Наприклад, слово «людина» має два основні значення і, відповідно, входить до складу двох окремих синсетів: "людина, людина розумна (вид тварин, що на сучасному етапі існування живого перебуває на найвищому щаблі розвитку і зайняв його в результаті довгого і складного процесу історико-еволюційного прогресу)" та "людина, особа (член суспільства)". Цілком зрозуміло, що кожне з цих значень займає своє місце в ієрархії (див. рис. 3, 4).

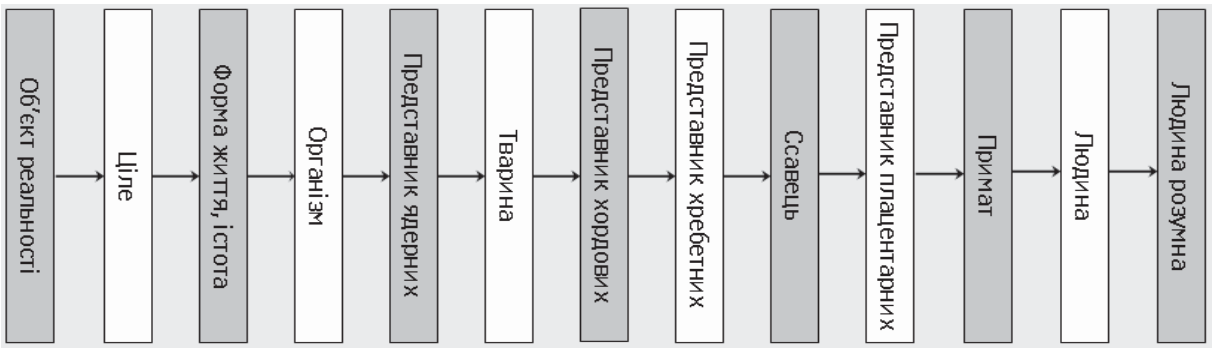

Рис. 3. Людина як біологічний вид у загальній ієрархії іменників

Рис. 4. Людина як член суспільства в загальній ієрархії іменників

Серед базових понять чимало професій та посад; до найуживаніших понять належать також почуття та емоції людини, риси характеру, споруди, величини (особливо, часові проміжки) тощо. Деякі ієрархії подано на рис. 5, 6, 7.

Рис. 5. Концепт особи, що займається певним видом професійної діяльності

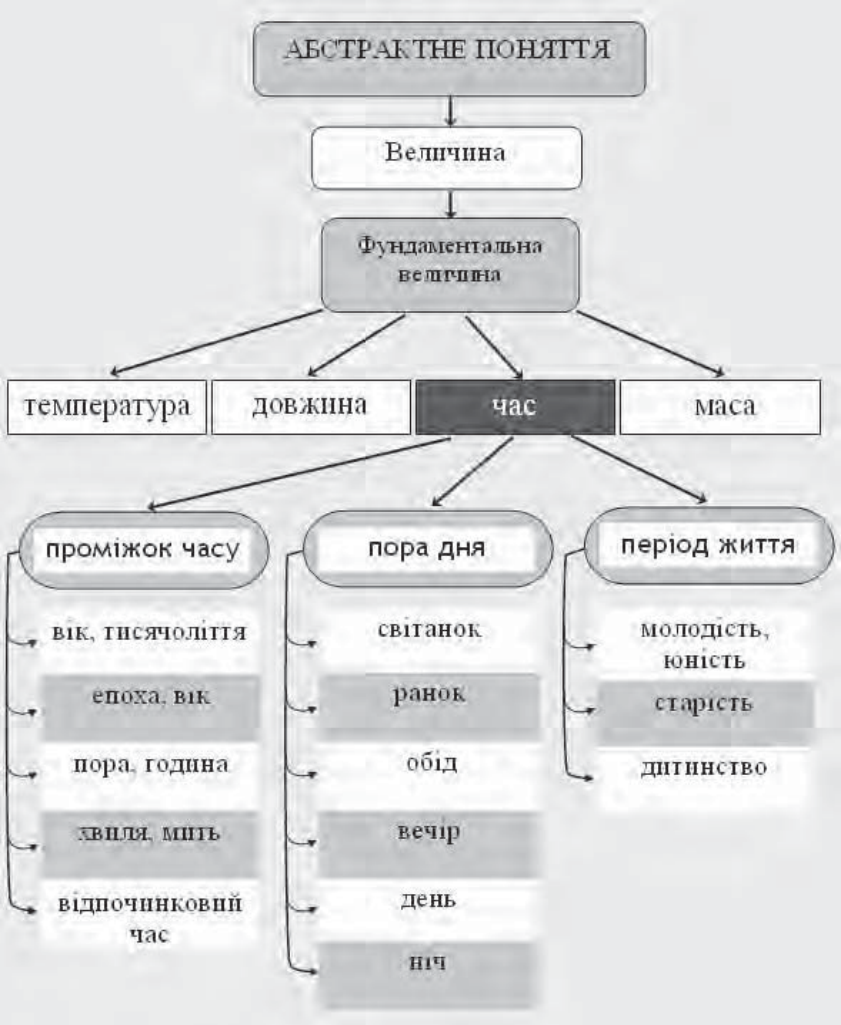

Рис. 6. Поняття часу в системі іменників

У випадку з емоціями та почуттями організація понять в українському WordNet-подібному словнику зовсім не відповідає ієрархії в англійському прототипі. Згідно з теоріями провідних вітчизняних і закордонних психологів, ані почуття, ані емоції не можна вважати ширшим поняттям. Вони відрізняються за тривалістю, постійністю та інтенсивністю вияву, за способом виникнення та універсальністю. Після аналізу низки праць про емотивні переживання людини було вирішено організувати іменники, що позначають різноманітні почуття та емоції, так, як зображено на рис. 7.

З наведених ієрархій видно, що саме родо-видове відношення забезпечує ефективну організацію іменників. Однак важливу роль відіграє також меронімія / голонімія. Так відношення частина-ціле має велике значення для утворення ієрархій біологічних видів, кожен із яких є представником відповідного таксону (одиниці біологічної класифікації). Якщо повернутися до «людини» як представника свого біологічного виду, побачимо складне переплетіння гіпо- / гіпеонімії та метонімії / голоінімії (див. рис. 8).

Цікаво, що спосіб організації іменників найвищого рівня ієрархії, характерний для версії WordNet 3.0. (рис. 1), зазнав кардинальних змін порівняно з попередніми версіями. У початковому вигляді всі іменники англійської мови погруповано так, що було виділено 25 так званих початкових понять, які не мали гіперонімів. Після подальшого аналізу їх кількість скоротилася до 11 (entity, abstraction, psychological feature, natural phenomenon, activity, event, group, location, possession, shape та state). Така систематизація до деякої міри виправдана, оскільки вона полегшує процес визначення місця того чи іншого слова в іменниковій ієрархії. Найновіший варіант ієрархії відображає намагання скоротити кількість вихідних понять та об'єднати всю систему іменників під одним загальним поняттям, однак таке зовнішнє спрощення насправді може бути причиною

виникнення додаткових труднощів. Для прикладу, якщо раніше поняття «власність» (possession), визначене як будь-який матеріальний або нематеріальний об'єкт, право розпоряджатися яким належить певній особі, займало своє чітке місце у загальній класифікації, то сьогодні, зважаючи на те, що всі поняття поділено на дві групи: об'єкт реальності (physical entity) й абстрактне поняття (abstraction), знайти йому належне місце взагалі неможливо.

Рис. 7. Емоції та почуття в загальній ієрархії

Крім цього, у результаті зменшення кількості вихідних понять, виникає небезпека, що структура лексикону набуде штучного характеру (це насамперед стосується похідних WordNet-подібних словників).

Досліджуючи організацію Прінстонського WordNet, можна знайти й інші протиріччя та розбіжності. Наприклад, поняття, виражене англійським словом activity (діяльність), до якого подано тлумачення «певний вид поведінки» вважається в найновішій версії гіперонімом поняття «поведінка». Отже, виникає суперечність між визначення та положенням в ієрархії. Інше поняття 'trait', тобто риса характеру, займає місце родового поняття щодо 'nature', 'character' та 'demeanour' (вдача, характер і ставлення, відповідно), хоча насправді воно є конкретним виявом риси, характеристики і складовою (тобто, меронімом) того самого характеру.

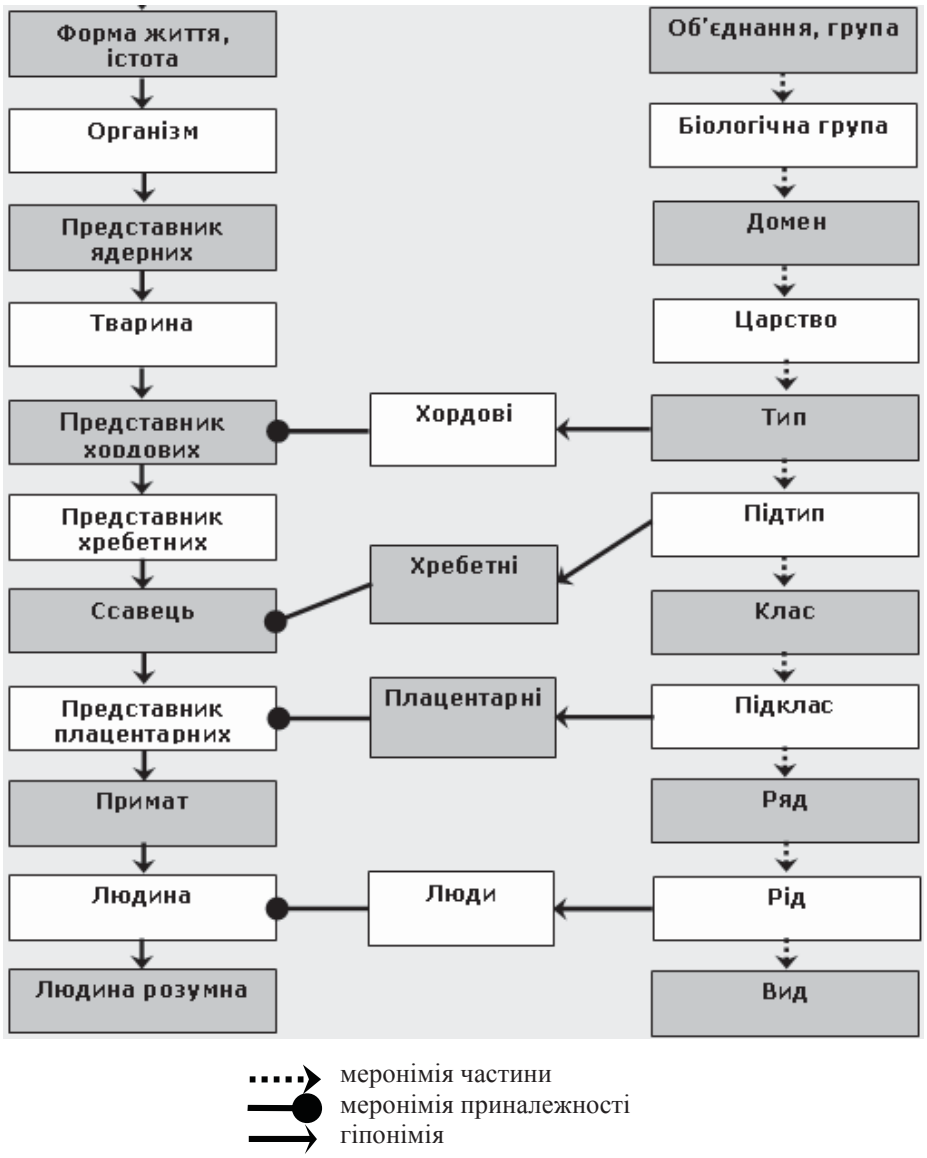

Рис. 8. Людина як біологічний вид в іменниковій ієрархії

Чимало питань викликає також підхід до виділення окремих значень слів у WordNet. Так звана «роздрібненість», «розпорошеність» значення доволі часто є предметом дискусій та суперечок дослідників. З одного боку, це перевага, оскільки в такий спосіб система забезпечує визначення точнішого значення, що є вкрай важливим для подолання багатозначності. Однак, з іншого боку, інколи зафіксовані у WordNet значення мають суто ситуативний характер або ж передбачають вживання у складі ідіоматичних одиниць. Перший випадок можна проілюструвати, проаналізувавши подання англійського слова ‘thing’ («річ» майже у всіх контекстах, «таке», «щось» у випадках, про які йтиметься далі). У версії WordNet 3.0. виділяють 12 окремих значень цієї лексичної одиниці, що зазвичай цілком обґрунтовано, зважаючи на розмитість семантики згаданого слова. Однак неоднозначно виглядає виокремлення значень «дія» та «висловлювання, що виступає предметом чогось» з огляду на майже ідентичні приклади вживання (“how could you do such a thing” – як ти міг / могла таке зробити; і “how can you say such a thing” – як ти можеш таке казати), адже поняття дії охоплює висловлювання.

Серед інших десяти тлумачень бачимо «артефакт», «подія», «специфічне узагальнення (абстрактне поняття)» тощо. Аналогічно, на основі прикладу вживання на зразок: “poor thing, it was a hard blow for her” (бідолашна, для неї це був справжній удар), можна розглядати ‘thing’ у значенні «особа», але таке значення у WordNet не зафіксовано.

Ще одним недоліком баз даних, побудованих за принципами Прінстонської моделі, є висока ймовірність суб'єктивності, оскільки організація понять та слів у створеному словнику певною мірою залежить від способу концептуалізації світу та розуміння понять, притаманних лексикографові-авторові.

Останнім часом багато уваги приділяють збагаченню WordNet переносними значеннями. Одна з найгостріших проблем, пов'язана з цим завданням, – визначення належного місця в ієрархії. Тоді як метонімія не викликає особливих труднощів, метафоричність лексичних одиниць досі залишається проблемою.

Підсумовуючи, можна зауважити, що Прінстонському WordNet усе ж притаманні деякі недоліки, неточності та суперечності. Проте це не впливає на його популярність та застосовність для вирішення численних завдань комп'ютерної лінгвістики. Постійно виникають нові проекти та нові галузі застосування, у яких така лексико-семантична база даних виявляється надзвичайно корисним інструментом. Робота над удосконаленням WordNet не припиняється, дослідники розробляють нові методи побудови, реалізовують нові типи відношень (напр., семантико-морфологічні, рольові та тематичні тощо) у національних словниках такої структури. Сьогодні існує понад 40 семантичних словників окремих мов, розроблених на основі WordNet, що свідчить про ефективність лексичних баз даних такої структури. Популярність прінстонського лексикону можна пояснити також доступністю публікацій про WordNet, та безкоштовним розповсюдженням самого словника, а також налагодженою міжнародною співпрацею.

6. Висновки

У результаті дослідження було розроблено фрагмент WordNet-подібного словника української мови, в якому реалізовано 194 синсети, пов'язані між собою зв'язками гіпо-/гіперонімії (183 приклади), антонімії (14 прикладів), а також додатково зв'язками меронімії/голонімії (понад 150 випадків) та ALSO_SEE (6 випадків).

Отримані результати дослідження дають змогу впевнено говорити про подальшу розбудову та збагачення WordNet, продовження зростання популярності цього ресурсу серед дослідників та необхідність продовження робіт із створення WordNet-подібного словника української мови.

1. Havasi, C. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge / Catherine Havasi, Robert Speer, Jason B. Alonso // Proceedings of Recent Advances in Natural Languges Processing, – Borovets, Bulgaria, 2007. 2. Chi-Yung Wang Knowledge-based Sense Pruning using the HowNet: an Alternative to Word Sense Disambiguation : A Thesis of Master of Philosophy In Computer Science / Wang Chi-Yung. – Hong Kong, 2002. – 111p. 3. Vanderwende L. MindNet: an automatically-created lexical resource / Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki, Arul Menezes // Proceedings of HLT/EMNLP Demonstration Abstracts, – Vancouver, 2005. – pp 8–9. 4. FrameNetII: Extended Theory and Practice [Електронний ресурс] / J. Ruppenhofer, M. Ellsworth, Miriam R.L. Petruck, Cristopher R.Johnson, Jan Scheffczyk // – 2006. – 166 pp. – Режим доступу: http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126. 5. Beth Levin. English Verb Classes and Alternations: A Preliminary Investigation / Beth Levin. – Chicago: The University of Chicago Press, 1993, – 348 pp. 6. Kipper K. Building a class-based verb lexicon using TAGs. /Karin Kipper, Hoa Trang Dang, William Schuler, Martha Palmer // TAG+5. Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms, – Paris, France, 2000. 7. WordNet: An Electronic Lexical Database / [Reem Al-Halimi, Robert C. Berwick, J. F. M. Burg etc.]; Edited by Christiane Fellbaum. – Cambridge, MA: MIT Press; 1998. – 422 pp. 8. Tufis D. Romanian WordNet: New Developments and Applications / D. Tufis, V. B. Mititelu, L. Bozianu, et al. // Proceedings of the Third International WordNet Conference. – Jeju Island, Korea, 2006. – pp. 337-344. 9. Корпусна лінгвістика / [Широков В.А., Булгаков О.В.,Грязнухина Т.О. та ін.]. – К. : Довіра, 2005. – 471 с. 10. Building a multilingual database with wordnets for several European languages. [Електронний ресурс]. / Режим доступу: http://www.illc.uva.nl/EuroWordNet/ 11. BALKANET: Design and Development of a Multilingual Balkan WordNet. [Електронний ресурс] / Режим доступу: http://www.ceid.upatras.gr/Balkanet/