УДК 811.161.2+811.111:004.9
DOI https://doi.org/10.32782/IT/2025-1-15

Наталія КОВАЛЕНКО
кандидат філологічних наук, доцент кафедри іноземних мов, філології та журналістики, Київський університет інтелектуальної власності та права Національного університету «Одеська юридична академія», вул. Харківське шосе, 210, м. Київ, Україна, 02121
***ORCID:** 0000-0001-9881-4835*

Уляна ЖОРНОКУЙ
кандидат філологічних наук, завідувач кафедри іноземних мов, філології та журналістики, Київський університет інтелектуальної власності та права Національного університету «Одеська юридична академія», вул. Харківське шосе, 210, м. Київ, Україна, 02121
***ORCID:** 0000-0001-9881-4835*
***Scopus Author ID:** 57260086300*

Бібліографічний опис статті: Коваленко, Н., Жорнокуй, У. (2025). Сучасні підходи та технології створення тезаурусів української та англійської мови: компаративний аналіз. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 105–113, doi: https://doi.org/10.32782/IT/2025-1-15

СУЧАСНІ ПІДХОДИ ТА ТЕХНОЛОГІЇ СТВОРЕННЯ ТЕЗАУРУСІВ УКРАЇНСЬКОЇ ТА АНГЛІЙСЬКОЇ МОВИ: КОМПАРАТИВНИЙ АНАЛІЗ

Актуальність. *Стрімкий розвиток інформаційних технологій та зростаюча роль штучного інтелекту в обробці природної мови створюють нові виклики та можливості у сфері створення тезаурусів. Особливої актуальності набуває порівняльне дослідження технологій створення тезаурусів для української та англійської мов, враховуючи їхні структурні відмінності та різний рівень розвитку відповідного інструментарію. В умовах глобалізації та діджиталізації якісні тезауруси стають незамінними для розвитку машинного перекладу, інформаційного пошуку та міжкультурної комунікації.*

Мета. *Здійснити порівняльний аналіз сучасних технологій створення тезаурусів української та англійської мов, визначити їхні спільні та відмінні риси, дослідити особливості застосування інструментів обробки природної мови та машинного навчання для кожної з мов, а також окреслити перспективи розвитку тезаурусної лексикографії в умовах цифрової епохи.*

Методологія. *У дослідженні застосовано комплексний підхід до аналізу технологій створення тезаурусів, що включає порівняльний, системний та структурний методи. Проведено детальний аналіз існуючих інструментів обробки природної мови, алгоритмів машинного навчання та корпусних менеджерів. Особлива увага приділена вивченню специфіки застосування технологій Word2Vec, BERT, NLTK, spaCy для української та англійської мов. Досліджено особливості роботи з корпусами текстів та семантичними мережами для обох мов.*

Наукова новизна. *Вперше проведено системний порівняльний аналіз технологічних особливостей створення тезаурусів української та англійської мов з урахуванням сучасних досягнень у сфері штучного інтелекту та обробки природної мови. Визначено специфічні виклики та обмеження у створенні україномовних тезаурусів, пов'язані з морфологічною складністю мови та обмеженістю доступних ресурсів.*

Висновки. *Дослідження показало, що створення сучасних тезаурусів вимагає поєднання традиційних лексикографічних методів із новітніми технологіями обробки природної мови та машинного навчання. Виявлено суттєві відмінності у доступності та розвиненості інструментарію для англійської та української мов, що впливає на процес створення тезаурусів. Визначено перспективні напрямки розвитку технологій для україномовних тезаурусів, включаючи створення спеціалізованих мовних моделей та розширення корпусів текстів. Обґрунтовано необхідність подальшого розвитку інструментів обробки природної мови з урахуванням специфіки української мови.*

Ключові слова: *тезаурус, обробка природної мови, машинне навчання, корпусна лінгвістика, семантичні мережі, українська мова, англійська мова.*

Nataliia KOVALENKO
Candidate of Philological Sciences, Associate Professor at the Department of Foreign Languages, Philology, and Journalism, Kyiv University of Intellectual Property and Law of the National University "Odesa Law Academy", 210, Kharkivske highway, Kyiv, Ukraine, 02121, reseacher.ds@gmail.com
***ORCID:** 0000-0001-9881-4835*

Uliana ZHORNOKUI
Candidate of Philological Sciences, Head of the Department of Foreign Languages, Philology, and Journalism, Kyiv University of Intellectual Property and Law of the National University "Odesa Law Academy", 210, Kharkivske highway, Kyiv, Ukraine, 02121, ulianazhornokui@gmail.com
ORCID: *0000-0001-9881-4835*
Scopus Author ID: *57260086300*

To cite this article: Kovalenko, N., Zhornokui, U. (2025). Suchasni pidkhody ta tekhnolohii stvorennia tezaurusiv ukrainskoi ta anhliiskoi movy: komparatyvnyi analiz [Modern Approaches and Technologies for Creating Thesauruses of the Ukrainian and English Languages: A Comparative Analysis]. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 105–113, doi: https://doi.org/10.32782/IT/2025-1-15

MODERN APPROACHES AND TECHNOLOGIES FOR CREATING THESAURUSES OF THE UKRAINIAN AND ENGLISH LANGUAGES: A COMPARATIVE ANALYSIS

Relevance. *The rapid development of information technologies and the growing role of artificial intelligence in natural language processing create new challenges and opportunities in thesaurus development. Comparative research of thesaurus creation technologies for Ukrainian and English languages becomes particularly relevant, considering their structural differences and varying levels of tool development. In the context of globalization and digitalization, quality thesauri have become indispensable for the development of machine translation, information retrieval, and intercultural communication.*

Objective. *To conduct a comparative analysis of modern technologies for creating thesauri of Ukrainian and English languages, identify their common and distinctive features, investigate the peculiarities of applying natural language processing tools and machine learning for each language, and outline the prospects for thesaurus lexicography development in the digital age.*

Methodology. *The study employs a comprehensive approach to analyzing thesaurus creation technologies, including comparative, systemic, and structural methods. A detailed analysis of existing natural language processing tools, machine learning algorithms, and corpus managers has been conducted. Special attention is paid to studying the specifics of applying Word2Vec, BERT, NLTK, and spaCy technologies for Ukrainian and English languages. The peculiarities of working with text corpora and semantic networks for both languages have been investigated.*

Scientific novelty. *For the first time, a systematic comparative analysis of technological features in creating thesauri for Ukrainian and English languages has been conducted, taking into account modern achievements in artificial intelligence and natural language processing. Specific challenges and limitations in creating Ukrainian-language thesauri, related to the morphological complexity of the language and limited available resources, have been identified.*

Conclusions. *The research has shown that creating modern thesauri requires combining traditional lexicographic methods with the latest natural language processing and machine learning technologies. Significant differences in the availability and development of tools for English and Ukrainian languages have been revealed, affecting the thesaurus creation process. Promising directions for developing technologies for Ukrainian-language thesauri have been identified, including the creation of specialized language models and expansion of text corpora. The necessity of further development of natural language processing tools considering the specifics of the Ukrainian language has been substantiated.*

Key words: *thesaurus, natural language processing, machine learning, corpus linguistics, semantic networks, Ukrainian language, English language.*

З огляду на стрімкий розвиток інформаційних технологій дослідження в галузі створення тезаурусів для української та англійської мов стають особливо нагальними. Така ситуація зумовлюється тим, що тезауруси відіграють ключову роль у розвитку штучного інтелекту, систем обробки й розуміння природної мови та автоматизованого перекладу. Вони є незамінними для систематизації та акумулювання знань, що є особливо важливим в умовах постійного збагачення словникового складу мов новими термінами та поняттями.

Актуальність теми підсилюється тим, що сучасна спільнота має велику потребу в систематизованій лінгвістичній інформації, яка відповідає світовим стандартам. Особливо важливим є створення як загальномовних, так і вузькогалузевих термінологічних тезаурусів, що сприяють розвитку міжнародних зв'язків, бізнесу, науки та культури. Порівняльне дослідження технологій створення тезаурусів української та англійської мов дозволяє не лише виявити унікальні характеристики кожної мови, але й визначити спільні тенденції в контексті глобальних лінгвістичних процесів.

Окрім того, розробка та впровадження сучасних технологій створення тезаурусів має практичне значення для освіти, адже дозволяє

створювати інтерактивні та адаптивні словники, що підвищують ефективність засвоєння знань. В умовах активної інтеграції України до світового простору дослідження технологій створення тезаурусів стає важливим інструментом збереження та розвитку української мови, одночасно сприяючи покращенню якості вивчення англійської мови для міжкультурної комунікації.

Аналіз досліджень. Дослідження сучасних підходів до створення тезаурусів активно розвиваються в Україні та світі завдяки зусиллям провідних вчених. Важливим аспектом є дослідження лексичного складу мови як основи для створення тезаурусів, чому присвятив свої дослідження О. О. Тараненко. Українські вчені О. В. Бісікало та О. В. Яхимович розробили практичні засоби автоматизації процесу ефективного створення тезаурусів. Г. Р. Мацюк аналізувала застосування тезаурусів у міждисциплінарних дослідженнях, тоді як А. Я. Гладун та Ю. В. Рогушина здійснили дослідження щодо використання онтологічного та мереологічного аналізу в контексті створення тезаурусів. Л. Л. Кармазіна розробляла методики побудови командних тезаурусів, а І. В. Асмукович досліджував принципи побудови двомовних тезаурусів. Таким чином, українські науковці охопили широкий спектр питань, пов'язаних зі створенням і застосуванням тезаурусів. Попри значні успіхи, існує низка проблем, що залишаються поза увагою дослідників, серед них можна згадати порівняльний аналіз підходів до створення тезаурусів в Україні та за кордоном, створення багатомовних тезаурусів, крос-культурна інтеграція тезаурусів, розробка нових методів автоматизації створення та оновлення тезаурусів

Мета статті. Здійснити порівняльний аналіз сучасних технологій створення тезаурусів української та англійської мов, визначити їхні спільні та відмінні риси, дослідити особливості застосування інструментів обробки природної мови та машинного навчання для кожної з мов, а також окреслити перспективи розвитку тезаурусної лексикографії в умовах цифрової епохи.

Виклад основного матеріалу. Тезаурус – це лексикографічний ресурс, який забезпечує систематизоване представлення слів і їхніх лексичних зв'язків, таких як синоніми, антоніми, гіпероніми та гіпоніми. Він спрямований на полегшення пошуку й використання мовних ресурсів у різноманітних контекстах. У цифрову епоху тезауруси стали не лише інструментом для лінгвістів, але й ключовим елементом у багатьох технологічних застосунках, таких як обробка природної мови, машинний переклад та пошукові системи.

Якщо проаналізувати дефініції поняття «тезаурус», то можна неозброєним оком помітити їхню різноплановість та варіативність. Таку ситуацію можна пояснити тим, що це поняття є міжгалузевим, адже ним активно послуговуються в таких різних науках як філософія, філологія, психологія, інформаційні технології, освіта, педагогіка тощо. Для наук гуманітарного спрямування в контексті створення та користування тезаурусами донедавна головною була функція структурованого накопичення знань, а для фахівців з інформаційних технологій важливо було забезпечити, окрім того, можливість швидкого пошуку інформації в базі даних, якою є для них тезаурус. Таким чином, визначення терміна тезаурус представниками відповідних галузей знань відображають ці загальногалузеві тенденції. Очевидно, що для повного розуміння змістовного наповнення цього терміна необхідно від рівня окремих визначень перейти на рівень виокремлення підходів до тлумачення терміна.

Прикладом такого підходу є стаття О. Тараненка в енциклопедії «Українська мова», де він розглядає тезаурус як «1) словник, що подає лексичний склад мови за семантичними розрядами (поняттєвими рубриками) з перехресним групуванням; 2) інформаційно-пошуковий словник, що подає в алфавітному порядку сукупність термінів (дескрипторів) певної галузі (галузей) знань із систематизацією їхніх ієрархічних та корелятивних відношень; 3) словник, завданням якого є повне охоплення лексичного складу мови» (Українська мова, 2004, с. 679). Як бачимо, перший та третій підходи є більш типовими для гуманітарного підходу, а другий підхід відображає тлумачення тезауруса в сфері інформаційних технологій.

Подальшим розвитком і осучасненням такого бачення є погляд Тур, яка на підставі узагальнення визначень терміна стверджує, що нині, по-перше, тезаурус треба розглядати як ідеографічний словник, в якому представлено слова мови та їхні семантичні відношення, по-друге, тезаурус треба сприймати як семантичну систему національної чи формалізованої мови для інформаційних пошукових систем; по-третє, тезаурус є складною системою, яка акумулює наші знання та уявлення про дійсність та метаінформацію, тобто він є чимось на кшталт метатезауруса. (Тур, 2014, с. 16)

Наразі будемо розглядати тезаурус як комплексну мовно-інформаційну систему, що визначає структуру знання про навколишній світ окремого індивіда й суспільства загалом у певному синхронному зрізі, забезпечує систематизацію,

зберігання, оновлення й пошук лексичних одиниць та семантичних зв'язків між ними. Таке визначення враховує традиційні лінгвістичні підходи до тезауруса як словника особливого типу та сучасне бачення його як динамічної складної інформаційно-пошукової системи.

Однією з можливих методологій для створення тезаурусів можна вважати методологію розробки онтологічних моделей, описану в праці О. Бісікала та О. Яхимовича. Вона передбачає виконання п'яти комлексних дій: 1) визначення і систематизація початкових умов; 2) збирання та накопичення даних; 3) аналіз даних; 4) початкова розробка тезауруса; 5) уточнення та затвердження тезауруса. Можна погодитись з тим, що ця схема цілком адекватно відображає загальну схему створення тезаурусів. (Бісікало, 2016, с. 30). Безперечно, що коли ми говоримо про створення тезаурусі у сучасному цифровому середовищі, то необхідно додати етап автоматизації та машинного навчання, на якому застосовуватимуться передусім алгоритми NLP, та машинного навчання для автоматичного виявлення семантичних зв'язків та розширення тезауруса. (Гладун, 2008).

Робота над тезаурусом потребує підбору та аналізу корпусу текстів, які відображають сучасний стан мови або репрезентують певну історичну епоху, якщо проводяться діахронічні дослідження. Корпуси можуть складатися з літературних творів, наукових статей, публіцистичних текстів та будь-яких інших джерел. Першочергове завдання дослідників на цьому етапі – забезпечити достатнє різноманіття текстів, щоб отримати представницький зріз мовних одиниць і контекстів їх використання. Наприклад, під час створення тезаурусів англійської мови використовуються корпуси British National Corpus (British National Corpus, 2024) або COCA (Corpus of Contemporary American English, 2024). Аналіз текстових корпусів включає визначення частотності слів, їхніх контекстів та семантичних зв'язків. На цьому етапі застосовуються методи статистичного аналізу, які дозволяють визначити найбільш уживані слова, фразеологізми та ідіоми.

Жодний тезаурус не може бути створений без застосування методів лінгвістичного аналізу. Серед таких методів у першу чергу треба назвати метод морфологічного аналізу, який забезпечує вивчення структури слів, їхніх форм та граматичних характеристик; метод синтаксичного аналізу, що дозволяє аналізувати структури речень для виявлення граматичних залежностей між словами та метод семантичного аналізу, який стає помічним при ідентифікації значень слів та їхніх семантичних зв'язків. Застосування сукупності цих методів забезпечує якісну обробку текстових даних, яка необхідна для створення точного та функціонального тезауруса.

Пришвидшити та автоматизувати лінгвістичний аналіз корпусів текстів допомагають алгоритми машинного навчання. Так, для виявлення лексичних зв'язків використовують моделі Word2Vec (Word2Vec, 2024) або GloVe (GloVe, 2024), які дозволяють представити слова у вигляді векторів, що відображають їх семантичну близькість, саме завдяки цьому стає можливим автоматичне формування груп синонімів, антонімів, а також виявлення прихованих зв'язків між словами. Отже, робота над тезаурусом на початковому етапі передбачає вибір корпусів текстів, методів лінгвістичного аналізу та алгоритмів обробки корпусів.

Структура тезауруса базується на різних типах семантичних зв'язків між лексичними одиницями, що дозволяє користувачам ефективно досліджувати семантичні поля та знаходити необхідну інформацію. Серед основних компонентів тезаурусу з лінгвістичної точки зору слід згадати лексичні одиниці та різні типи зв'язків між ними (Кармазіна, 2023). Такі лексичні одиниці як синоніми є ключовим компонентом тезаурусу, адже вони забезпечуючи варіативність висловлювання та точне передавання значень. Сучасні тезауруси враховують не лише семантичну близькість, але й стилістичні відмінності та контексти вживання синонімів. Антоніми ж відображають протилежні значення слів, сприяючи глибшому розумінню семантичної структури мови та організації лексики за принципом опозиції. У свою чергу, гіпероніми (родові поняття) та гіпоніми (видові поняття) формують ієрархічну структуру лексики, відображають відношення загального до конкретного. Важливо також при створенні тезаурусів враховувати поняття меронімії (відношення «частина-ціле»), наприклад, «колесо» є меронімом до «автомобіль».

Базовими типами зв'язків у тезаурусі є тематичні, які об'єднують слова за спільністю тематики або сфери використання, та асоціативні, які відображають слова, що часто вживаються разом або викликають спільні асоціації. Тематична організація полегшує навігацію по тезаурусу та пошук потрібної лексики, а асоціативні зв'язки важливі для розуміння конотативного значення слів та створення інтуїтивно зрозумілого інтерфейсу, особливо в електронних тезаурусах. Окрім того, вони також використовуються в системах рекомендацій та пошуку інформації.

Створення сучасних тезаурусів є складним процесом, що поєднує традиційні лексикографічні методи з сучасними комп'ютерними технологіями. Розглянемо основні технології та інструменти для створення тезаурусів. Одним із ключових інструментів для створення сучасних тезаурусів є програмне забезпечення для аналізу корпусів текстів, адже аналіз корпусів текстів є фундаментальним для створення тезаурусів. Воно дозволяє збирати, обробляти та аналізувати великі обсяги текстових даних, необхідних для побудови якісного тезаурусу.

Сучасні фахівці з цією метою використовують Sketch Engine (Sketch Engine, 2024), що являє собою програмне забезпечення для роботи з корпусами текстів і є потужним інструментом. Воно дозволяє виявляти частотність слів, створювати списки словосполучень, аналізувати їхні контексти та будувати дистрибутивну семантику. Однією з основних переваг Sketch Engine є можливість працювати з багатомовними корпусами, що робить його ідеальним інструментом для створення тезаурусів для різних мов. Цей інструмент підтримує функцію автоматизованого аналізу, яка значно пришвидшує процес формування списків синонімів, антонімів і тематичних зв'язків. Наприклад, використовуючи Sketch Engine, можна легко визначити, які слова найчастіше вживаються в контексті певного терміну, що допомагає встановити їх асоціативні зв'язки.

Сучасні технології обробки природної мови (NLP) є основою для створення електронних тезаурусів. Вони використовуються для автоматизації процесу аналізу текстів, виявлення семантичних зв'язків між словами та їхньої класифікації. Для роботи з тезаурусами найчастіше використовують наступні бібліотеки:

• NLTK (Natural Language Toolkit) – це бібліотека Python з широким набором інструментів для токенізації, морфологічного аналізу, синтаксичного аналізу, стемінгу, лематизації та інших NLP-задач. NLTK надає можливість аналізувати великі текстові корпуси, що дозволяє визначати частотність слів, їхні граматичні характеристики та тематичні зв'язки (Natural Language Toolkit, 2024).

• spaCy – це сучасна та ефективна бібліотека Python для обробки природної мови з підтримкою багатомовних моделей, що забезпечує швидкий та точний аналіз текстів, включно із розпізнавання іменованих сутностей, синтаксичним аналізом та векторизацією слів. Її ключовою перевагою є здатність інтегруватися з іншими інструментами машинного навчання для глибшого аналізу текстів (spaCy, 2024).

Використання алгоритмів машинного навчання для створення тезаурусів є одним із найперспективніших напрямів сучасної лінгвістики. Завдяки цим технологіям можна автоматизувати процес ідентифікації лексичних зв'язків і класифікації слів. Машинне навчання також використовується для кластеризації лексичних одиниць. Наприклад, алгоритми кластеризації дозволяють групувати слова за тематикою або стилістичними ознаками, що є важливим для створення спеціалізованих тезаурусів.

• Word2Vec – ця модель створює векторні представлення слів, які відображають їхні семантичні зв'язки. Наприклад, слова, що використовуються в схожих контекстах, мають близькі вектори у багатовимірному просторі. Це дозволяє виявляти синоніми, антоніми та інші асоціативні зв'язки (Word2Vec, 2024).

• GloVe – алгоритм, що комбінує статистичний підхід і методи машинного навчання для визначення співвідношень між словами. Він ефективний для створення багатомовних тезаурусів, оскільки дозволяє визначати спільні семантичні поля для різних мов (GloVe, 2024).

• інші методи машинного навчання, такі як кластеризація (k-means, DBSCAN), класифікація (SVM, Random Forest) та нейронні мережі, також використовуються для створення та розширення тезаурусів.

Семантичні мережі є потужним інструментом для моделювання складних лексичних зв'язків. Вони дозволяють структурувати інформацію у вигляді графів, де вузли представляють слова, а ребра – їхні зв'язки. WordNet є однією з найвідоміших семантичних мереж для англійської мови. Вона містить інформацію про синоніми, антоніми, гіпероніми, гіпоніми та інші зв'язки. WordNet використовується як основа для створення тезаурусів у багатьох мовах. (Мацюк, 2019). Семантичні мережі забезпечують інтуїтивно зрозуміле представлення даних, що дозволяє користувачам легко знаходити потрібні слова та їхні зв'язки. Крім того, вони можуть бути інтегровані в системи штучного інтелекту, що забезпечує автоматизацію пошуку інформації.

Для управління великими обсягами текстових даних використовуються спеціалізовані корпусні менеджери. Вони дозволяють ефективно організовувати, зберігати та аналізувати текстові корпуси. Corpus Workbench (CWB) – це інструмент для управління текстовими базами даних, що підтримує багатомовний аналіз. CWB забезпечує швидкий доступ до великих обсягів текстових даних та інструменти для їхнього синтаксичного й семантичного аналізу (Corpus Workbench, 2024).

Аналіз великих обсягів текстових даних є важливим елементом створення сучасних тезаурусів. Використання технологій великих даних дозволяє обробляти мільйони текстів, що забезпечує репрезентативність і точність створених лексичних ресурсів.

• Hadoop – це фреймворк для розподіленої обробки великих даних, що дозволяє зберігати та обробляти петабайти інформації на кластері комп'ютерів, платформа для зберігання й обробки великих обсягів даних, яка дозволяє працювати з текстовими корпусами в розподілених системах (Hadoop, 2024).

• Spark – це швидкий та потужний фреймворк для обробки даних у реальному часі, що часто використовується для аналізу текстів та машинного навчання, сучасний інструмент для аналізу даних у реальному часі. Spark використовується для швидкого оброблення текстових даних та створення лексичних моделей (Spark, 2024).

Попри спільність підходів до створення тезаурусів української та англійської мов існує чимало аспектів, які треба враховувати під час роботи над тезаурусами. Ключовою є відмінність в лексичних, морфологічних, синтаксичних та культурних особливостях. Розглянемо розбіжності в цих мовах за трьома основними аспектами: зміст та структура, процеси та процедури створення, інструменти та ресурси.

До категорії зміст та структура будемо відносити відмінності у лексичних одиницях, морфології, синтаксисі, фразеології та ідіомах. Лексичні одиниці відрізняються тим, що англійська мова відома великою кількістю синонімів, часто з тонкими відмінностями у значенні та стилістичному забарвленні. Наприклад, для слова «small» існує багато синонімів, таких як «tiny», «petite», «little», «micro», кожен з яких має свої конотації. В українській мові синонімічні ряди часто менші за кількістю, але відмінності між синонімами можуть бути більш вираженими та залежними від контексту.

Складна флективна система української мови з розгалуженою системою відмінювання, роду, числа та інших граматичних категорій обумовлює морфологічні відмінності та створює значні труднощі для автоматичної обробки та створення тезаурусів. В українській мові дуже часто визначення синонімів залежить від граматичних форм слова. Англійська мова має аналітичну структуру та відносно просту морфологію, що значно полегшує автоматизацію цих процесів.

Різниця в синтаксичній структурі мов також впливає на формування тезаурусів. Українська мова допускає більш вільний порядок слів у реченні, що ускладнює автоматичний аналіз семантичних зв'язків. В англійській мові з фіксованим порядком слів цей аналіз є значно простішим.

Ідіоми та фразеологізми, зазвичай, відображають культурні особливості мови, однак, не всі ідіоми мають такі прямі відповідники, тому це створює труднощі при створенні міжмовних тезаурусів.

В контексті відмінностей в процесах та процедурах створення тезаурусів української та англійської мов треба констатувати, що для англійської мови існує велика кількість великих цифрових корпусів текстів, таких як British National Corpus (BNC), Corpus of Contemporary American English (COCA), Google Books Ngrams (Google Books Ngrams, 2024), що містять мільярди слів і забезпечують репрезентативну базу для аналізу. Українська мова ще не має настільки потужних цифрових корпусів текстів. Серед найбільших можна назвати Генеральний регіонально анотований корпус української мови (ГРАК), його колекція містить тексти з 1816 по 2022 рік і охоплює понад 130 тисяч різножанрових текстів; Корпус української мови, що має 120 млн. слововживань; Лабораторія Української (веб-корпус із синтаксичною розміткою) з 3 млрд. Токенів; UkTenTen: Ukrainian corpus from the Web, який нараховує 7,5 млрд. токенів і містить тексти з інтернету (ГРАК, 2024; Корпус української мови, 2024; UkTenTen, 2024; Лабораторія Української, 2024).

Щодо процедури аналізу текстів, то треба зазначити, що для української мови алгоритми синтаксичного та морфологічного аналізу є складнішими через її складну морфологію та синтаксис української мови, ніж для англійської. Програмні засоби та моделі повинні враховувати велику кількість відмінкових закінчень, родових форм та інші морфологічні особливості.

Моделі машинного навчання, такі як Word2Vec, GloVe, BERT, RoBERTa, широко використовуються для аналізу англійської мови та створення векторних представлень слів (Olizarenko, 2024; Kumari, 2023). Розробка та застосування подібних моделей для української мови є більш складним завданням через меншу кількість доступних даних та складнішу структуру мови, однак вже існують успішні, але мало використовувані приклади цього, такі як ukrBERT. UkrBERT – модель машинного навчання, яка базується на архітектурі BERT (Bidirectional encoder representations from transformers) і адаптована для української мови (Ukrainian News Classification Experiments,

2022). Вона була створена для покращення точності обробки текстів українською мовою, її використовують для різноманітних завдань обробки текстів на кшталт класифікації текстів, розпізнавання мовлення, генерація тексту.

Для роботи з україномовними текстами можна також задіювати багатомовні моделі. Як приклад можна назвати XLM-R (Cross-lingual Language Model – RoBERTa), яка є мультикультурною моделлю, яка підтримує багато мов, включно з українською, вона використовується для багатомовних завдань і може допомогти в класифікації текстів, розпізнаванні мовлення та інших задачах (XLM-RoBERTa, 2024). Модель mBERT (Multilingual BERT) також є багатомовною моделлю, це багатомовна версія BERT, яка підтримує багато мов, серед них і українську (mBERT, 2024).

Вже з викладеного вище стає зрозуміло, що для створення тезаурусів українською мовою є значно менше ресурсів та інструментів. Багато інструментів для обробки текстів, такі як NLTK, spaCy, Sketch Engine, спочатку були розроблені для англійської мови, тому їх треба було адаптовувати для української мови, а це вимагало додаткових зусиль та створення спеціальних мовних моделей.

Подібна ситуація склалася і щодо лінгвістичних ресурсів для створення тезаурусів. Англійська мова має такий універсальний лінгвістичний ресурс як WordNet, який значно спрощує створення тезаурусів, натомість для української мови створення подібних ресурсів є нагальним завданням, хоч вона і має деякі електронні словники, які можна використати як джерело тезаурусів. Слід констатувати, що наявні інструменти та ресурси в першу чергу розробляються для англійської мови, а вже потім адаптуються до українських лінгвістичних реалій, починають враховувати специфічну граматику, лексику та культурний контекст.

Висновки і перспективи подальших досліджень. Створення тезаурусів у цифрову епоху є комплексним процесом, що поєднує традиційні лексикографічні методи із сучасними комп'ютерними технологіями та вимагає врахування специфіки конкретної мови. Основні відмінності у створенні тезаурусів української та англійської мов зумовлені: морфологічними особливостями (складна флективна система української мови проти аналітичної структури англійської) різним обсягом доступних цифрових корпусів; різним рівнем розвитку інструментарію для обробки текстів; культурно-специфічними особливостями лексики. Для англійської мови існує більше розвинених інструментів та ресурсів (WordNet, великі корпуси текстів, адаптовані NLP-інструменти), тоді як для української мови багато технологічних рішень перебувають на етапі розробки або адаптації. Перспективним напрямком є розвиток специфічних для української мови інструментів обробки текстів та створення великих мовних корпусів, що дозволить покращити якість створюваних тезаурусів. Використання сучасних технологій машинного навчання (Word2Vec, BERT, тощо) та їх адаптація для української мови відкриває нові можливості для автоматизації процесу створення тезаурусів. Розвиток тезаурусів обох мов має важливе значення для розвитку штучного інтелекту, систем машинного перекладу та міжкультурної комунікації.

ЛІТЕРАТУРА:

1. Українська мова: енциклопедія / НАН України, Ін-т мовознав. ім. О. О. Потебні, Ін-т укр. мови; редкол.: В. М. Русанівський [та ін.]. Вид. 2-ге, випр. і допов. Київ : Вид-во «Українська енциклопедія» ім. М. П. Бажана, 2004. 820 с.

2. Тур О. Лексикографічні витоки сучасних концепцій тезаурусного моделювання термінолексики. *Вісник Книжкової палати*. 2014. № 8. С. 15–17.

3. Бісікало О. В., Яхимович О. В. Автоматизоване визначення лексичних онтологій з тезаурусу технічного спрямування. *Оптико-електронні інформаційно-енергетичні технології*. 2016. № 1. С. 26–38.

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

5. British National Corpus http://www.natcorp.ox.ac.uk/ (дата звернення: 21.12.2024).

6. Corpus of Contemporary American English. URL: https://www.english-corpora.org/coca/ (дата звернення: 21.12.2024).

7. Word2Vec. URL: https://www.tensorflow.org/text/tutorials/word2vec (дата звернення: 19.12.2024).

8. GloVe. URL: https://nlp.stanford.edu/projects/glove/ (дата звернення: 19.12.2024).

9. Кармазіна Л. Л. Методологія розробки та принципи побудови командного тезаурусу. *Закарпатські філологічні студії*. 2023. Т. 1, вип. 27. С. 198–202.

10. Sketch Engine. URL: https://www.sketchengine.eu/ (дата звернення: 15.12.2024).

11. Natural Language Toolkit. URL: https://www.nltk.org/ (дата звернення: 15.12.2024).

12. spaCy. URL: https://spacy.io/ (дата звернення: 13.12.2024).
13. Мацюк Г. Р. Інформаційно-пошукові тезауруси: світовий та вітчизняний досвід формування. *Бібліотекознавство. Документознавство. Інформологія*. 2019. № 2. С. 106–115.
14. Corpus Workbench. URL: https://cwb.sourceforge.io/ (дата звернення: 14.12.2024).
15. Hadoop. URL: https://hadoop.apache.org/ (дата звернення: 17.12.2024).
16. Spark. URL: https://spark.apache.org/ (дата звернення: 11.12.2024).
17. Google Books Ngrams. URL: https://books.google.com/ngrams/ (дата звернення: 19.12.2024).
18. Генеральний регіонально анотований корпус української мови (ГРАК) URL: uacorpus.org (дата звернення: 20.12.2024).
19. Корпус української мови. URL: http://www.mova.info/corpus.aspx?l1=209 (дата звернення: 21.12.2024).
20. UkTenTen: Ukrainian corpus from the Web. URL: https://www.sketchengine.eu/uktenten-ukrainian-corpus/ (дата звернення: 20.12.2024).
21. Лабораторія Української (веб-корпус із синтаксичною розміткою). URL: https://www.sketchengine.eu/uktenten-ukrainian-corpus/ (дата звернення: 19.12.2024).
22. Olizarenko S., Argunov V. On possibilities of multilingual BERTmodel for determining semantic similarities of the news content. *Системи управління, навігації та зв'язку*. 2020. Вип. 3 (61). С. 94–98.
23. Kumari K. RoBERTa: A Modified BERT Model for NLP. URL: https://www.comet.com/site/blog/roberta-a-modified-bert-model-for-nlp/ (дата звернення: 10.12.2024).
24. Ukrainian News Classification Experiments. URL: https://github.com/StepanTita/news-contest (дата звернення: 12.12.2024).
25. XLM-RoBERTa. URL: https://huggingface.co/docs/transformers/ model_doc/xlm-roberta (дата звернення: 15.12.2024).
26. BERT multilingual base model (cased) URL: https://huggingface.co/google-bert/bert-base-multilingual-cased (дата звернення:17.12.2024).

REFERENCES:

1. Rusanivskyi, M. (Eds.). (2004). Ukrainska mova: entsyklopediia [Ukrainian Language: Encyclopedia] / NAN Ukrainy, In-t movoznav. im. O. O. Potebni, In-t ukr. movy; redkol.: V. M. Rusanivskyi [ta in.]. Vyd. 2-he, vypr. i dopov. Kyiv : Vyd-vo "Ukrainska entsyklopediia" im. M. P. Bazhana [in Ukrainian].
2. Tur, O. (2014). Leksykohrafichni vytoky suchasnykh kontseptsii tezaurusnoho modeliuvannia terminoleksyky [Lexicographic Origins of Modern Concepts of Thesaurus Modeling of Terminology]. *Visnyk Knyzhkovoi palaty*, 8, 15–17 [in Ukrainian].
3. Bisikalo, O. V., Yakhymovych, O. V. (2016). Avtomatyzovane vyznachennia leksychnykh ontolohii z tezaurusu tekhnichnoho spriamuvannia [Automated Definition of Lexical Ontologies from Technical Thesaurus]. *Optyko-elektronni informatsiino-enerhetychni tekhnolohii*, 1, 26–38 [in Ukrainian].
4. Hladun, A. Ya, Rohushyna, Yu. V. (2008). Osnovy metodolohii formuvannia tezaurusiv z vykorystanniam ontolohichnoho ta mereolohichnoho analizu [Basic Methodology of Thesaurus Formation Using Ontological and Mereological Analysis]. *Shtuchnyi intelekt*, 4, 53–61 [in Ukrainian].
5. British National Corpus. Retrieved from: http://www.natcorp.ox.ac.uk/ (date of access: 21.12.2024).
6. Corpus of Contemporary American English. Retrieved from: https://www.english-corpora.org/coca/ (date of access: 21.12.2024).
7. Word2Vec. Retrieved from: https://www.tensorflow.org/text/ tutorials/word2vec (date of access: 19.12.2024).
8. GloVe. Retrieved from: https://nlp.stanford.edu/projects/glove/ (date of access: 19.12.2024).
9. Karmazina, L. L. (2023). Metodolohiia rozrobky ta pryntsypy pobudovy komandnoho tezaurusu [Methodology of Development and Principles of Command Thesaurus Construction]. Zakarpatski filolohichni studii, 1, 27, 198–202 [in Ukrainian].
10. Sketch Engine. Retrieved from: https://www.sketchengine.eu/ (date of access 15.12.2024).
11. Natural Language Toolkit. Retrieved from: https://www.nltk.org/ (date of access: 15.12.2024).
12. spaCy. Retrieved from: https://spacy.io/ (date of access: 13.12.2024).
13. Matsiuk, H. R. (2019). Informatsiino-poshukovi tezaurusy: svitovyi ta vitchyznianyi dosvid formuvannia [Information Retrieval Thesauri: World and Domestic Experience of Formation]. *Bibliotekoznavstvo. Dokumentoznavstvo. Informolohiia*, 2, 106–115 [in Ukrainian].
14. Corpus Workbench. Retrieved from: https://cwb.sourceforge.io/ (date of access: 14.12.2024).
15. Hadoop. Retrieved from: https://hadoop.apache.org/ (date of access: 17.12.2024).

16. Spark. Retrieved from: https://spark.apache.org/ (date of access: 11.12.2024).

17. Google Books Ngrams. Retrieved from: https://books.google.com/ngrams/ (дата звернення: 19.12.2024).

18. Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK) [General Regionally Annotated Corpus of Ukrainian Language (GRAC)]. Retrieved from: uacorpus.org (date of access) [in Ukrainian].

19. Korpus ukrainskoi movy [Corpus of the Ukrainian Language]. Retrieved from: http://www.mova.info/corpus.aspx?l1=209 (date of access: 21.12.2024) [in Ukrainian].

20. UkTenTen: Ukrainian corpus from the Web. Retrieved from: https://www.sketchengine.eu/uktenten-ukrainian-corpus/ (date of access: 20.12.2024).

21. Laboratoriia Ukrainskoi (veb-korpus iz syntaksychnoiu rozmitkoiu) [Laboratory of the Ukrainian (Web-corpus with Syntactic Markup)]. Retrieved from: https://www.sketchengine.eu/uktenten-ukrainian-corpus/ (date of access: 19.12.2024) [in Ukrainian].

22. Olizarenko, S., Argunov, V. (2020). On possibilities of multilingual BERTmodel for determining semantic similarities of the news content. Systemy upravlinnia, navihatsii ta zviazku, 3 (61), 94–98.

23. Kumari K. RoBERTa: A Modified BERT Model for NLP. Retrieved from: https://www.comet.com/site/blog/roberta-a-modified-bert-model-for-nlp/ (date of access: 10.12.2024).

24. Ukrainian News Classification Experiments. Retrieved from: https://github.com/StepanTita/news-contest (date of access: 12.12.2024).

25. XLM-RoBERTa. Retrieved from: https://huggingface.co/docs/transformers/ model_doc/xlm-roberta (date of access: 15.12.2024).

26. BERT multilingual base model (cased) Retrieved from: https://huggingface.co/google-bert/bert-base-multilingual-cased (date of access:17.12.2024).