УДК 81’322:004.8:004.912
DOI https://doi.org/10.24919/2308-4863/90-2-33

Орест ТОРСЬКИЙ,
orcid.org/0009-0002-1896-1021
студент ІІ курсу магістратури
Інституту комп’ютерних наук та інформаційних технологій
Національного університету «Львівська політехніка»
(Львів, Україна) orestorskyi@gmail.com

Наталія КУНАНЕЦЬ,
orcid.org/0000-0003-3007-2462
професор кафедри інформаційних систем та мереж
Інституту комп’ютерних наук та інформаційних технологій
Національного університету «Львівська політехніка»
(Львів, Україна) nek.lviv@gmail.com

АКТУАЛЬНІСТЬ КОРПУСІВ ДАНИХ З РОЗВИТКОМ ШТУЧНОГО ІНТЕЛЕКТУ

У статті висвітлено актуальність використання корпусів даних у контексті стрімкого розвитку технологій штучного інтелекту (ШІ) та цифрової трансформації різних галузей знання. Корпуси даних розглядаються як базова інфраструктура для навчання, тестування, перевірки та вдосконалення алгоритмів машинного навчання й мовних моделей, які сьогодні широко застосовуються в комп’ютерній лінгвістиці, юриспруденції, освіті, перекладознавстві, кібербезпеці, медицині та інших сферах. Особливу увагу приділено проблемам якості, репрезентативності, відкритості й етичності корпусів, а також викликам, пов’язаним із нестачею україномовних ресурсів, їх нерівномірною структурованістю та технічною складністю використання.

У статті здійснено класифікацію корпусів за форматом даних (текстові, аудіо, відео, мультимодальні), типами (загальні, спеціалізовані, історичні, паралельні, мікро- та навчальні корпуси) та функціональними можливостями. Детально проаналізовано такі сучасні проєкти, як Cambridge Law Corpus, Forestry English Corpus і Common Corpus, які демонструють потенціал спеціалізованих ресурсів у розвитку міждисциплінарних рішень. Розглянуто інноваційні методи збору та обробки корпусних даних, включно з веб-скрейпінгом, псевдо-розміткою, семіавтоматичною індексацією та генеративним моделюванням, а також перспективи поєднання корпусних ресурсів із генеративними нейронними мережами, зокрема LLM-моделями.

Окрема увага приділена критичному дискурс-аналізу, освітнім технологіям, моделі DDL (Data-Driven Learning) і новітній концепції MRU (Metacognitive Resource Use), які демонструють синергію між людським аналітичним мисленням і машинною обробкою даних. У роботі також окреслено шляхи захисту вразливих мовних ресурсів за допомогою інструментів штучного інтелекту, зокрема систем кібербезпеки, блокчейн-контролю доступу та аномалієвого виявлення, що є важливими для збереження нематеріальної культурної спадщини.

Узагальнені результати демонструють, що корпуси даних не лише забезпечують мовні моделі ресурсами, а й формують нові зв’язки між лінгвістикою, інформаційними технологіями та соціогуманітарними науками. Подальше розширення, оптимізація та етичне управління корпусними ресурсами є необхідною умовою розвитку адаптивних та соціально значущих технологій штучного інтелекту.

Ключові слова: *корпуси даних, штучний інтелект, мовні моделі, DDL, MRU, генеративні технології, мультимодальність, етичність, дискурс-аналіз, освітні технології.*

Orest TORSKYI,
orcid.org/0009-0002-1896-1021
2nd year master's student
Institute of Computer Science and Information Technologies of Lviv Polytechnic National University
(Lviv, Ukraine) orestorskyi@gmail.com

Nataliia KUNANETS,
orcid.org/0000-0003-3007-2462
Professor at the Department of Information Systems and Networks
Institute of Computer Science and Information Technologies of Lviv Polytechnic National University
(Lviv, Ukraine) nek.lviv@gmail.com

THE RELEVANCE OF DATA CORPORA WITH THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE

The article highlights the relevance of using data corpora in the context of the rapid development of artificial intelligence (AI) technologies and the digital transformation of various fields of knowledge. Data corpora are considered a fundamental infrastructure for training, testing, evaluating, and improving machine learning algorithms and language models, which are now widely used in computational linguistics, law, education, translation studies, cybersecurity, medicine, and beyond. Particular attention is paid to issues of corpus quality, representativeness, openness, and ethical use, as well as challenges related to the scarcity of Ukrainian-language resources, their obsolescence, structural inconsistency, and technical complexity of implementation.

The article presents a classification of corpora by data format (textual, audio, video, multimodal), by type (general, specialized, historical, parallel, micro- and learner corpora), and by functional applications. A detailed analysis is provided of recent corpus projects such as the Cambridge Law Corpus, Forestry English Corpus, and Common Corpus, which demonstrate the potential of domain-specific resources for interdisciplinary innovation. The study examines modern methods of data collection and processing, including web scraping, pseudo-annotation, semi-automatic indexing, and generative modeling, as well as the integration of corpora with generative neural networks, particularly large language models (LLMs).

Special attention is given to critical discourse analysis, educational technologies, the Data-Driven Learning (DDL) model, and the emerging Metacognitive Resource Use (MRU) framework, which illustrate the synergy between human analytical thinking and machine-based data processing. The paper also outlines approaches to safeguarding endangered linguistic resources using AI-based tools such as cybersecurity systems, blockchain-based access control, and anomaly detection, which are vital for preserving intangible cultural heritage.

The findings demonstrate that corpora serve not only as language model resources but also as a foundation for a new paradigm of interaction between linguistics, information technology, and the social sciences. The further expansion, optimization, and ethical governance of corpus resources are presented as essential conditions for the development of transparent, adaptive, and socially impactful AI technologies.

Key words: *data corpora, artificial intelligence, language models, DDL, MRU, generative technologies, multimodality, ethics, discourse analysis, educational technologies.*

Постановка проблеми. Корпуси даних є основою для розвитку сучасних технологій штучного інтелекту (ШІ). У теперішніх умовах цифровізації та автоматизації зростає потреба у великих, репрезентативних та високоякісних наборах даних, які використовуються для навчання, тестування та вдосконалення інтелектуальних систем. Саме дані визначають ефективність, точність та безпечність моделей, які застосовують у різних галузях.

Проте сьогодні існує низка проблем, які стримують розвиток ШІ через недостатність якісних корпусів. По-перше, це обмежена кількість корпусів українською мовою та для специфічних галузей. Як наслідок, створення конкурентних мовних та спеціалізованих моделей суттєво ускладнюється. По-друге, самі дані часто виявляються неповними, нерепрезентативними або застарілими – це формує викривлені результати роботи моделей та їхню упередженість. По-третє, відсутність належної системи відкритого доступу до корпусів даних гальмує наукові дослідження та практичну розробку прикладних рішень. Так, у сукупності ці фактори значно обмежують прогрес у розвитку штучного інтелекту.

Водночас, розвиток генеративного ШІ, комп'ютерного зору, систем автоматичного перекладу та багатьох інших напрямів безпосередньо залежить від наявності великих і різноманітних корпусів, які охоплюють сучасну лексику, розмовні варіанти мови, професійну термінологію та реальні контексти застосування. Таким чином, можна стверджувати, що усвідомлення важливості корпусів даних як ефективного ресурсу для розвитку національних технологій ШІ є необхідним.

Аналіз досліджень. У дослідженнях розвитку штучного інтелекту (ШІ) корпуси даних розглядаються як базовий ресурс, який забезпечує навчання, тестування та оцінювання моделей у реальних умовах (Zappavigna, 2023).

У науковій літературі *корпуси даних* визначаються як впорядковані мовні або мультимодальні зібрання текстів, які створюються за певними принципами для подальшого аналізу, досліджень або навчання моделей (Boulton, Tyne, 2013). Корпус може містити тисячі або мільйони одиниць даних, що робить його надійною базою для дослідження мовних закономірностей, створення словників, граматик та тренування алгоритмів.

У дослідженнях розвитку штучного інтелекту (ШІ) корпуси даних розглядаються як базовий ресурс, який забезпечує навчання, тестування та оцінювання моделей у реальних умовах (Incelli, 2025). У науковій літературі поняття корпус даних також трактується як впорядкований набір структурованих або неструктурованих даних, що використовується для тренування та перевірки алгоритмів машинного навчання й мовних моделей у різних галузях, включаючи лінгвістику, комп'ютерний зір, юриспруденцію та медицину (Paquot, Gries, 2021).

Залежно від завдань, корпуси можуть містити текстові документи, аудіозаписи, відео або мультимодальні дані, що дає можливість забезпечувати моделі максимальною кількістю реальних прикладів для аналізу та генерації (Hamat, Heryanto, 2019).

Наприклад, створення спеціалізованих корпусів, таких як Cambridge Law Corpus, який включає понад 250 тисяч судових рішень Великобританії, сприяє розробці та вдосконаленню моделей ШІ для автоматичного витягання юридичних висновків та прогнозування результатів справ. Це значно підвищує ефективність правових досліджень та практичної діяльності юристів (Östling et al., 2023), Такі приклади підтверджують, що корпуси даних є не лише джерелом мовної інформації, а й потужним інструментом для розвитку міждисциплінарних ШІ-систем.

Водночас розвиток генеративного ШІ та його інтеграція у корпусні дослідження викликають активні дискусії серед науковців. Хоча моделі на зразок ChatGPT здатні швидко семантично категоризувати ключові слова та проводити базовий аналіз, результати їхньої роботи часто є поверхневими або недостатньо точними для глибинного аналізу спеціалізованих дискурсів, особливо коли йдеться про складні функціонально-комунікативні категорії (Curry et al., 2024).

Дослідження також підкреслюють, що поєднання корпусної лінгвістики та ШІ відкриває нові перспективи, зокрема у створенні мовних помічників, удосконаленні систем автоматичного перекладу та розпізнавання мовлення, а також при розробці навчальних платформ з адаптивним контентом (Pace-Sigge, 2018). У цих галузях корпуси даних забезпечують моделі не лише лексичним та синтаксичним матеріалом, а й допомагають виявляти приховані семантичні зв'язки, стилістичні особливості та дискурсивні структури, що є критично важливим для генерації природної мови.

Крім того, сучасні підходи до створення корпусів ґрунтуються на використанні технологій вебскрейпінгу, автоматичної розмітки та індексації з застосуванням алгоритмів машинного навчання, що значно прискорює процес збирання великих обсягів даних. Проте це, у свою чергу, ставить нові завдання перед дослідниками та розробниками щодо забезпечення точності, достовірності, релевантності та етичності таких ресурсів, особливо в умовах зростаючої кількості автоматично згенерованого контенту у відкритих джерелах (Hong, 2018).

Метою статті є аналіз сучасного стану та значення корпусів даних у контексті розвитку штучного інтелекту, зокрема їхньої ролі у тренуванні, тестуванні та вдосконаленні мовних моделей. Особливу увагу приділено класифікації корпусів за видами даних та сферами застосування, а також висвітленню проблем та викликів, пов'язаних із забезпеченням якості та репрезентативності корпусів у добу розвитку генеративного ШІ. Завданням роботи є узагальнення основних підходів до створення корпусів, визначення їхньої актуальності для розвитку національних та міждисциплінарних технологій штучного інтелекту, а також окреслення перспектив подальших досліджень у напрямку розширення та оптимізації корпусних ресурсів.

Виклад основного матеріалу. У сучасних лінгвістичних та міждисциплінарних дослідженнях корпуси даних відіграють ключову роль як базові ресурси для аналізу, моделювання, навчання та тестування алгоритмів. Під *корпусами* зазвичай розуміють впорядковані масиви автентичних текстових, аудіо, відео або мультимодальних даних, призначених для дослідження мови в реальному вжитку, а також для тренування моделей штучного інтелекту у різних сферах (Pérez-Paredes, Ordoñana-Guillamón, 2025). Існують численні класифікації корпусів; одна з них передбачає поділ на general, specialised, parallel, historical та multimodal corpora (Murphy, Riordan, 2016).

Варто зазначити, що універсального стандарту класифікації корпусів не існує. Проте у науковій практиці сформувалися певні підходи, які дозволяють систематизувати корпуси за їхньою структурою, змістом, мовними характеристиками, функціональним призначенням та технічною організацією (Жуковська, 2013). Так, у сучасній лінгвістиці наголошується на існуванні численних моделей класифікації, які відображають різні дослідницькі пріоритети та завдання (Коцюк, Коцюк, 2020). На основі узагальнення підходів, запропонованих у працях Жуковської (2013) та Коцюк Л. і Коцюка І. (2020), можна виокремити такі типи класифікації корпусів (Жуковська, 2013; Коцюк Л., Коцюк І., 2020):

1. **За мовною складовою**:

– *Мономовні (монолінгвальні)* – містять тексти однією мовою (наприклад, British National Corpus).

– *Багатомовні (мультилінгвальні)* – включають тексти кількома мовами без прямого зіставлення.

– *Паралельні корпуси* – вирівняні за реченнями або сегментами тексти в оригіналі та перекладі (напр., Europarl).

– *Компарабельні корпуси* – тематично подібні тексти різними мовами, що не мають відповідності на рівні речень.

2. **За жанром і стилем**:

– літературні,

– публіцистичні,

– наукові,

– розмовні (транскрибовані усні виступи),

– інтернет-комунікація (пости, коментарі, форуми),

– спеціалізовані (медичні, технічні, правничі тощо).

3. **За типом дискурсу**:

– *монологічні* (доповіді, статті),

– *діалогічні* (чати, інтерв'ю),

– *полілогічні* (групові обговорення, форуми).

4. **За рівнем розмітки**:

– *неанотовані* (сирі тексти),

– *лінгвістично анотовані* (морфологічна, синтаксична, семантична, прагматична розмітка),

– *з параметрами/метаданими* (інформація про автора, дату, жанр тощо),

– *з онтологічною або дискурсивною розміткою* (тематика, комунікативна мета, соціальні ролі тощо).

5. **За способом формування**:

– *балансовані* – репрезентативні щодо певної мови або жанру,

– *спеціалізовані* – сформовані для конкретної тематики або домену,

– *динамічні* – оновлюються в реальному часі (наприклад, корпуси соцмереж),

– *ручного збирання* або *автоматичні* (з використанням веб-скрейпінгу та ML-алгоритмів).

6. **За обсягом**:

– *малі* (до 1 млн токенів),

– *середні* (1–100 млн),

– *великі* (100+ млн),

– *мега- та гіга-корпуси* (мільярди токенів – напр., Common Crawl).

7. **За доступністю**:

– *відкриті* (open access),

– *комерційні або ліцензовані*,

– *закриті або внутрішні* (використовуються у корпоративних або закритих дослідженнях).

8. **За функціональним призначенням**:

– для лінгвістичних і дискурс-аналізів,

– для тренування NLP-моделей і алгоритмів машинного навчання,

– для лексикографії,

– для досліджень у перекладознавстві,

– для навчання LLM (Large Language Models),

– для освітніх цілей і моделі DDL.

Щоб краще проілюструвати специфіку зазначених критеріїв класифікації, було обрано низку найвідоміших лінгвістичних корпусів, які відрізняються за мовним складом, структурою, розміром, ступенем розміченості та функціональним призначенням. Для зручності сприйняття їхні характеристики зведено в узагальнювальну таблицю, яка дозволяє виявити типові та унікальні риси для кожного корпусу.

Загалом дані для лінгвістичних досліджень надходять або з інтуїтивних уявлень про мову, або з спостережень за мовленнєвими подіями. Саме зібрання таких даних у вигляді корпусів дозволило у XX ст. здійснити прорив у лінгвістиці завдяки електронному зберіганню та пошуку великих обсягів текстів, що замінило залежність лише від інтуїції дослідників (Aarts, 2022). General corpora охоплюють широке коло жанрів та стилів, забезпечуючи репрезентативність мови загалом. Їхні обсяги зростають до трильйонів слів, як у COCA, і вони дедалі частіше включають тексти з цифрової комунікації (наприклад, блоги); так, це розширює спектр їх використання в педагогіці, особливо у DDL (Pérez-Paredes, Ordoñana-Guillamón, 2025). Specialised corpora зосереджуються на конкретних галузях – юридичній, медичній, науковій – і потребують чіткого маркування контекстуальних характеристик для правильної інтерпретації. Їх створення відіграє важливу роль у формуванні професійної лексики. Наприклад, Forestry English Corpus продемонстрував підвищення ефектив-

Таблиця 1

Приклади класифікації реальних корпусів за декількома критеріями.
Examples of classification of real corpuses by multiple criteria

Назва корпусу	Мовна складова	Жанрова структура	Рівень розмітки	Обсяг	Призначення
COCA (https://www.english-corpora.org/coca/)	Англійська (монолінгвальний)	Публіцистика, художня література, телемовлення, академічні, розмовні	Морфологічна, синтаксична (лінгвістично анотований)	1+ млрд токенів	Дослідження мовних змін, лексикографія, навчання
GRAC (https://uacorpus.org/)	Українська (монолінгвальний)	Публіцистика, література, офіційні тексти	Морфологічна, частково синтаксична	500+ млн токенів	Лінгвістика, лексикографія
British National Corpus (https://www.english-corpora.org/bnc/)	Англійська (монолінгвальний)	Широкий спектр жанрів	Морфологічна	~ 100 млн токенів	Лінгвістичний аналіз, лексикографія
Sketch Engine Corpus (https://www.sketchengine.eu/)	Багатомовний	Веб-тексти	Морфологічна, синтаксична	1+ млрд токенів	Лексикографія, термінологія
The International Corpus of English (ICE) (https://www.ice-corpora.uzh.ch/en/design.html)	Англійська (монолінгвальний)	Академічна мова, офіційні виступи, діалоги, розмови, ЗМІ	Морфологічна, синтаксична, структурна	~1 млн токенів	Порівняльна лінгвістика, варіативність англійської
The NOW corpus (News on the Web) (https://www.english-corpora.org/now/)	Англійська (монолінгвальний)	Новини, онлайн-журнали	Морфологічна, синтаксична	22.6 млрд токенів (2025); +3.1 млрд щороку	Моніторинг змін у сучасній англійській, вивчення нових слів і трендів
Common Crawl (https://commoncrawl.org/)	Мультилінгвальний (50+ мов, домінує англійська)	Різножанрові веб-тексти (новини, блоги, форуми, наукові сайти, комерційні сторінки)	Неанотовані (сирі тексти у HTML та WARC-форматі) у вихідній версії; мають параметри/метадані (дата збору, URL, MIME-тип, розмір файл)	≈400–500 млрд токенів у повному архіві; обсяг зростає щомісяця	Створення великих мовних моделей (LLM), багатомовних корпусів, тренування пошукових та класифікаційних систем
Prague Dependency Treebank (PDT-C 1.0) (https://lindat.mff.cuni.cz/repository/home)	Чеська (монолінгвальний)	Газети, WSJ-переклади, діалоги, UGC	Морфологічна, синтаксична	~1,5 мільйона токенів	NLP: синтаксичний парсинг, лінгвістичні дослідження
CLASSLA-web (https://www.clarin.si/info/k-centre/)	Південнослов'янські мови (Мультилінгвальний)	Великий набір веб-жанрів, позначені жанри за допомогою класифікатора X-GENRE	Автоматична лінгвістично анотована розмітка	13 млрд токенів із 26 млн документів	Порівняння слов'янських мов, аналіз жанрів, лінг. дослідження
Europarl Corpus (https://www.statmt.org/europarl/)	Паралельний	Стенограми засідань Європарламенту (офіційні документи)	Лінгвістично анотована + з парметрами / метаданими	+ 1,2 млрд токенів для 21 мови	Машинний переклад, багатомовні дослідження

ності засвоєння термінології студентами лісогосподарських спеціальностей на 27,3% завдяки інтеграції корпусів у навчання (Fang, 2023).

Parallel corpora містять тексти мовою оригіналу та їх переклади іншою мовою. Вони використовуються в перекладознавстві та контрастивній лінгвістиці, адже дозволяють досліджувати відмінності синтаксичних і морфологічних конструкцій різних мов, проте мають обмеження репрезентативності через домінування художньої літератури, парламентських дебатів і технічних мануалів (Lefer, 2021). Розширення мовної й жанрової бази є перспективним напрямом їх розвитку, адже це дозволить підвищити якість автоматичного перекладу та розробку міжмовних моделей ШІ.

Historical corpora використовуються для дослідження мовної еволюції. Їхніми пріоритетами є створення методологій для дослідження як довготривалих, так і нещодавніх змін, забезпечення сумісності корпусів між собою, а також глибоке дослідження соціокультурного (Lindemann, Alonso Arrospide, 2024)

Multimodal corpora поєднують текстові, аудіо-, візуальні та поведінкові дані, що дає можливість досліджувати взаємодію мовних конструкцій з іншими семіотичними ресурсами, наприклад жестами чи інтонацією, та реалізовувати DDL у поєднанні з невербальними аспектами комунікації (Coccetta, 2022). Проте їхнє створення вимагає значних технічних ресурсів та специфічного програмного забезпечення для анотування й синхронізації різних каналів даних. Learner corpora, що містять мову учнів (як носіїв, так і неносіїв), дозволяють досліджувати типові помилки та специфіку міжмовної інтерференції, що є важливим для створення ефективних освітніх технологій. Водночас micro-corpora або ad hoc corpora, які формуються для вузьких навчальних чи дослідницьких цілей, хоч і мають невеликі обсяги, проте забезпечують високу релевантність матеріалу для конкретної аудиторії (Mörth et al., 2011).

Новітнім явищем є Common Corpus – найбільший відкритий корпус для тренування LLM, що налічує понад два трильйони токенів багатьма мовами світу та забезпечує відповідність етичним і правовим вимогам, оскільки містить лише дані без авторських обмежень або під вільними ліцензіями [8]. Це особливо важливо з огляду на сучасні регуляторні обмеження у використанні даних для тренування ШІ (Liang et al., 2022). У юридичній сфері створюються спеціалізовані legal corpora, які використовують семіавтоматичні ML-фреймворки, що поєднують апроцедури псевдо-розмітки, GAN-генерації даних та ансамблеве навчання для досягнення високої повноти, актуальності й коректності (Chen et al., 2022). Це демонструє поєднання корпусної лінгвістики з сучасними AI-практиками.

Варто зазначити, що класифікація корпусів здійснюється не лише за функціональними ознаками (загальні, спеціалізовані, паралельні тощо), а й за критеріями цільового використання, структури, способу збирання та анотації даних. У межах розвитку штучного інтелекту все більшої популярності набувають data-centric підходи, які передбачають класифікацію корпусів за ступенем чистоти даних, збалансованістю категорій, рівнем анотації та типом завдань, що вирішуються: класифікація, генерація, прогнозування тощо. Зокрема, широко використовуються instruction-tuned corpora, які включають пари запит–відповідь для навчання моделей типу GPT, а також reinforcement datasets, що дозволяють оптимізувати поведінку моделі в умовах людської оцінки (RLHF). Такі корпуси супроводжуються багаторівневою семантичною та прагматичною розміткою, що критично важливо для забезпечення контекстно обґрунтованих відповідей ШІ-моделей у складних завданнях.

Корпуси також активно використовуються у машинному перекладі. Так, у Китаї створено численні паралельні корпуси, що лягли в основу розвитку інтелектуального перекладу та сценарно-орієнтованих систем, зокрема шляхом інтеграції open-source data, crowdsourcing translation, self-learning models та human-machine cooperation (Hou, 2022). Такі моделі забезпечують масштабованість, швидкість оновлення й адаптивність корпусів до різних доменів та мовних стилів. У свою чергу, використання корпусів у генерації автоматичних відповідей чат-ботами передбачає застосування методів NLP, ML-класифікації та глибинних нейронних мереж, що дозволяє створювати адекватні та швидкі відповіді на запити користувачів, що є критично важливим у розвитку інтелектуальних помічників та сервісів підтримки клієнтів (Коваленко, Сердюк, 2020).

Таким чином, корпуси даних у сучасних лінгвістичних, освітніх та міждисциплінарних дослідженнях не лише виступають джерелом мовної інформації, а й є основою розвитку технологій штучного інтелекту, генеративних моделей, спеціалізованих інтелектуальних систем, правових фреймворків та освітніх практик DDL.

Проте, поряд із цими можливостями, дедалі більшої уваги потребує й якість корпусного контенту. Адже мовні дані, які використовуються для тренування AI-систем, не завжди є нейтральними та збалансованими. Як зазначають дослідники (Jones, 2024), корпуси, які лягають в основу генеративних мовних AI-систем, можуть містити упередження; це зумовлює ризик упередженого мовлення в автоматизованих відповідях. Так, можна стверджувати, що це особливо критично в освітньому контексті, де штучний інтелект використовується як інструмент навчання. Для мінімізації цього ризику пропонують реалізацію етичного підходу до добору даних, верифікація джерел і залучення багатомовних та мультикультурних корпусів.

У ширшому міждисциплінарному контекстівсе більше уваги приділяється також збереженню вразливих мовних ресурсів, зокрема мов, які зникають. Так, дослідження Suba Linguistic Corpus демонструє інтеграцію засобів штучного інте-

лекту й кібербезпеки задля захисту культурно значущих мовних даних від втрат та деяких викривлень (Ondiba, 2025). Тому, було запропоновано поєднання аномалієвого виявлення, блокчейн-контролю доступу та принципів етичного управління даними. Такий підхід показує потенціал AI не лише як мовного інструменту, а й як засобу для збереження нематеріальної культурної спадщини.

Грунтуючись на попереніх твердженнях, можна зробити висновок, що корпусно орієнтоване навчання набуває ще більшої ефективності при інтеграції зі штучним інтелектом. Наприклад, у дослідженні англомовного ESP-письма студентів традиційної китайської медицини показано, що поєднання DDL-моделі з AI-аналізаторами сприяє підвищенню якості письма за п'ятьма ключовими критеріями, зокрема структури тексту та мотивації (Su et al., 2024). Така синергія дозволяє студентам самостійно досліджувати мовний матеріал і формувати глибше розуміння професійної лексики.

Синергетичне поєднання корпусних методів і ШІ також відкриває нові перспективи в аналізі соціально важливих дискурсів. Корпуси дедалі активніше застосовуються у критичному дискурс-аналізі (CDA), зокрема для вивчення політичних, медійних та технологічних наративів. Так, порівняльне дослідження державних політик щодо штучного інтелекту в Північній Америці та країнах Східної Азії, здійснене методом CDA на основі спеціально сформованих корпусів, виявило чіткі регіональні відмінності в акцентах, семантичному забарвленні та лексичному оформленні стратегічних документів (Gu, 2023). Подібні дослідження демонструють потенціал корпусної лінгвістики як інструменту виявлення прихованих ідеологем, риторичних стратегій та соціокультурних пріоритетів, що формують публічний і технологічний дискурс.

Таким чином, інтеграція корпусних ресурсів і засобів штучного інтелекту дедалі більше утверджується як потужна парадигма у сфері мовної освіти та аналітики дискурсу. Синергія цих інструментів дозволяє не лише оптимізувати навчальні процеси, а й активізувати глибші когнітивні стратегії у студентів, а також досліджувати соціокультурні процеси крізь призму мови. Особливого значення така інтеграція набуває в умовах переходу від традиційної DDL-моделі (Data-Driven Learning) до більш гнучких, мультиресурсних підходів.

Зокрема, новітні освітні концепції, як-от MRU (Metacognitive Resource Use), передбачають стратегічне комбінування корпусів із генеративними моделями ШІ. Це дозволяє студентам не лише самостійно працювати з автентичними мовними даними, а й розвивати метакогнітивну рефлексію, що є ключовим чинником у формуванні академічної автономності (Mizumoto, 2023). Інструменти на кшталт ChatGPT, у такому випадку, не замінюють, а доповнюють корпуси, створюючи динамічне освітнє середовище, яке реагує на запити й потреби користувача в режимі реального часу. Це свідчить про поступову трансформацію DDL у змішану модель навчання, орієнтовану на самостійну аналітичну діяльність.

У ширшому контексті застосування штучного інтелекту корпуси відіграють фундаментальну роль у розвитку текстової аналітики (text analytics) – міждисциплінарної галузі, яка об'єднує лінгвістику, інформатику та соціальні науки. Вона передбачає автоматизоване вилучення знань з великих масивів тексту, розпізнавання іменованих сутностей, тематичне моделювання та аналіз латентних семантичних структур. У такому підході корпуси виступають не лише джерелом лінгвістичних даних, а й ключовою ланкою у побудові моделей прогнозування, класифікації та виявлення тенденцій у суспільному дискурсі (Moreno, Redondo, 2023). Отже, мовні ресурси виконують чимало функцій та допомагають краще зрозуміти суспільні явища, відображаючи, як мова реагує на зміни у світі. Завдяки цьому корпуси иожна вважати корисним та потужним інструментом для осмислення глобальних процесів через аналіз мовлення та дискурсу.

Висновки. Узагальнюючи викладене, можна стверджувати, що мовні корпуси залишаються ключовим інструментом у сучасній лінгвістиці, а їхнє поєднання з можливостями штучного інтелекту значно розширює спектр застосування. Інтеграція корпусів у навчання, особливо в межах концепцій DDL та MRU, сприяє розвитку автономного мислення та глибшого засвоєння мовного матеріалу. У дослідницькому контексті корпусні дані є основою для критичного аналізу дискурсів, вивчення професійної комунікації та ідеологічно маркованих наративів.

Окрему увагу приділено етичним і технічним аспектам: потребі у верифікованих, відкритих та мультикультурних корпусах, а також важливості захисту мов, що зникають. Використання інструментів кібербезпеки, блокчейн-технологій і аномалієвого виявлення ілюструє новий рівень відповідальності у сфері лінгвістичних досліджень.

Таким чином, корпуси в поєднанні з генеративними моделями та аналітичними системами перетворюються на потужний інструмент не лише навчання й дослідження, а й осмислення мовної картини сучасного світу. У цьому контексті мовні ресурси постають важливою ланкою між лінгвістикою, технологіями та гуманітарним знанням загалом.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Zappavigna Michele. Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data. *Applied Corpus Linguistics*, 2023, 3.3: 100067. URL: https://doi.org/10.1016/j.acorp.2023.100067
2. Boulton Alex; Tyne, Henry. Corpus linguistics and data-driven learning: A critical overview. *Bulletin suisse de Linguistique appliquée*, 2013, 97: 97-118.
3. Incelli Ersilia. Exploring the Future of Corpus LInguistics: Innovations in AI and Social Impact. *International Journal of Mass Communication*, 2025, 3: 1-10. DOI: 10.6000/2818-3401.2025.03.01
4. Paquot M., Gries, S. T. A practical handbook of corpus linguistics. – Cham: Springer Nature, 2021. DOI: 10.1007/978-3-030-46216-1
5. Hamat Afendi, Heryanto Anton. Annotation tools for ai analysis of corpus data. Prosiding simposium kemanusiaan digital *2019*, 7. URL: https://www.researchgate.net/publication/336736801_PROSIDING_SIMPOSIUM_KEMANUSIAAN_DIGITAL_2019_Proceeding_of_Digital_Humanities_Symposium_2019-_COMPUTATIONAL_THINKING_FOR_FUTURE_DATA_SCIENTISTS_eISBN_978-883-9391-71-8
6. Östling Andreas, et al. The Cambridge law corpus: A dataset for legal AI research. *Advances in Neural Information Processing Systems*, 2023, 36: 41355-41385. URL: https://doi.org/10.48550/arXiv.2309.12269
7. Curry Niall, Baker Paul, Brookes Gavin. Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 2024, 4.1: 100082. URL: https://doi.org/10.1016/j.acorp.2023.100082
8. Pace-Sigge Michael. Where corpus linguistics and artificial intelligence (AI) meet. In: *Spreading Activation, Lexical Priming and the Semantic Web: Early Psycholinguistic Theories, Corpus Linguistics and AI Applications*. Cham: Springer International Publishing, 2018. p. 29-82. DOI: 10.1007/978-3-319-90719-2_3
9. Hong Changchun. Construction of corpus in Artificial Intelligence age. In: *MATEC Web of Conferences*. EDP Sciences, 2018. p. 03037. URL: https://doi.org/10.1051/matecconf/201817503037
10. Pérez-Paredes Pascual, Ordoñana-guillamón Carlos. Types of Corpora in Data-Driven Learning. In: *The Palgrave Encyclopedia of Computer-Assisted Language Learning*. Cham: Springer Nature Switzerland, 2025. p. 1-10. DOI: 10.1007/978-3-031-51447-0_68-1
11. Murphy Bróna, Riordan Elaine. Corpus types and uses. In: *The Routledge handbook of language learning and technology*. Routledge, 2016. p. 388-403. DOI: 10.4324/9781315657899
12. Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник. 2013. URL: https://eprints.zu.edu.ua/18909/1/korpusna_lingv.pdf
13. Коцюк Л. М., Коцюк, Ю. А. Класифікаційна парадигма корпусу текстів за особли-востями його дизайну, структури та способами використання, а також способом фіксації та індексації текстових даних (Classificational paradigm of a text corpus by its design, structure and use, as well as by the fixation and indexation methods of its text data). *Наукові записки Національного університету «Острозька академія». Серія «Філологія»*, 2020, 9 (77): 106-110. URL: http://eprints.oa.edu.ua/id/eprint/9556
14. Aarts Jan. Corpus analysis. In: *Handbook of Pragmatics*. John Benjamins Publishing Company, 2022. p. 1505-1515. DOI: 10.1075/hop.m2.cor2
15. Fang Mingyuan. Forestry english corpus construction and application in foreign language teaching under the background of big data and internet of things. *Journal of Computational Methods in Science and Engineering*, 2023, 23.1: 211-221. URL: https://doi.org/10.3233/JCM-226562
16. Lefer Marie-Aude. Parallel corpora. In: *A practical handbook of corpus linguistics*. Cham: Springer International Publishing, 2021. p. 257-282. DOI: 10.1007/978-3-030-46216-1_12
17. Lindemann D. Alonso, Arrospide M. Linking historical corpus data and annotations using Wikibase. Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. – Zagreb: Institute for the Croatian Language, 2024. – P. 743–748. URL: https://doi.org/10.5281/zenodo.13932191
18. Coccetta Francesca. Multimodal corpora and concordancing in DDL. *The Routledge handbook of corpora and English language teaching and learning*, 2022, 361-376. URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003002901-29/multimodal-corpora-concordancing-ddl-francesca-coccetta
19. Mörth Karlheinz, Dorostkar Niku, Preisinger Alexander. Gleaning micro-corpora from the internet: integrating heterogeneous data into existing corpus infrastructures. *Editores María Luisa Carrió Pastor Miguel Ángel Candel Mora*, 2011, 111.
20. Liang Weixin, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 2022, 4.8: 669-677. URL: https://doi.org/10.1038/s42256-022-00516-1
21. Chen Haihua, Pieptea Lavinia F., Ding Junhua. Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches. *IEEE Transactions on Reliability*, 2022, 71.2: 657-673. DOI: 10.1109/TR.2022.3156126
22. Hou Xiaoqing. Research on translation corpus building with the assistance of ai. In: *2021 International conference on Smart Technologies and Systems for Internet of Things (STS-IOT 2021)*. Atlantis Press, 2022. p. 136-140. DOI: 10.2991/ahis.k.220601.027
23. Коваленко О. С., Сердюк О. А. Використання класичних методів машинного навчання для класифікації текстів у програмах генерації автоматичних відповідей. *Вісник Черкаського університету. Серія «Прикладна математика. Інформатика»*, 2020, 1: 86-100. DOI: 10.31651/2076-5886-2020-1-86-100
24. Jones John W. AI, Language, and Education: Bias and Prejudice in Text and Corpus Analysis. In: *Encyclopedia of Educational Innovation*. Springer, Singapore, 2024. p. 1-5. URL: https://doi.org/10.1007/978-981-13-2262-4_312-1

25. Ondiba Hesborn. Proactive AI-Driven Cybersecurity for Endangered Language Preservation: Safeguarding the Suba Linguistic Corpus. In: *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*. IEEE, 2025. p. 1-5. DOI: 10.1109/ICAIC63015.2025.10848675.

26. Su Zheqian, et al. Influence of "Corpus Data Driven Learning+ Learning Driven Data" Mode on ESP Writing Under the Background of Artificial Intelligence. In: *International Conference on Application of Intelligent Systems in Multi-modal Information Analytics*. Cham: Springer International Publishing, 2020. p. 547-552. URL: https://doi.org/10.1007/978-3-030-51431-0_80

27. Gu Feng. Corpus-based critical discourse analysis on AI policy: a comparison between North America and developing countries in East Asia. *Asian J Soc Sci Stud*, 2023, 8.3: 14. URL: https://doi.org/10.20849/ajsss.v8i3.1371

28. Mizumoto Atsushi. Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics*, 2023, 3.3: 100074. URL: https://doi.org/10.1016/j.acorp.2023.100074

29. Moreno Antonio, Redondo Teófilo. Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 2016, 3.6: 57-64. DOI: 10.9781/ijimai.2016.369

REFERENCES

1. Zappavigna Michele (2023) Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data. *Applied Corpus Linguistics*, 3.3: 100067. URL: https://doi.org/10.1016/j.acorp.2023.100067

2. Boulton Alex, Tyne Henry (2013) Corpus linguistics and data-driven learning: A critical overview. *Bulletin suisse de Linguistique appliquée*, 97: 97-118.

3. Incelli Ersilia, et al (2025) Exploring the Future of Corpus LInguistics: Innovations in AI and Social Impact. *International Journal of Mass Communication*, 3: 1-10. DOI: 10.6000/2818-3401.2025.03.01

4. Paquot M., Gries S. T., eds. (2021) A practical handbook of corpus linguistics. – Cham: Springer Nature. DOI: 10.1007/978-3-030-46216-1

5. Hamat Afendi, Heryanto Anton (2019) Annotation tools for ai analysis of corpus data. Prosiding simposium kemanusiaan digital, 7. URL: https://www.researchgate.net/publication/336736801_PROSIDING_SIMPOSIUM_KEMANUSIAAN_DIGITAL_2019_Proceeding_of_Digital_Humanities_Symposium_2019-_COMPUTATIONAL_THINKING_FOR_FUTURE_DATA_SCIENTISTS_eISBN_978-883-9391-71-8

6. Östling Andreas, et al (2023) The Cambridge law corpus: A dataset for legal AI research. *Advances in Neural Information Processing Systems*, 36: 41355-41385. URL: https://doi.org/10.48550/arXiv.2309.12269

7. Curry Niall, Baker Paul, Brookes Gavin (2024) Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 4.1: 100082. URL: https://doi.org/10.1016/j.acorp.2023.100082

8. Pace-Sigge Michael (2018) Where corpus linguistics and artificial intelligence (AI) meet. In: *Spreading Activation, Lexical Priming and the Semantic Web: Early Psycholinguistic Theories, Corpus Linguistics and AI Applications*. Cham: Springer International Publishing. p. 29-82. DOI: 10.1007/978-3-319-90719-2_3

9. Hong Changchun (2018) Construction of corpus in Artificial Intelligence age. In: *MATEC Web of Conferences*. EDP Sciences. p. 03037. URL: https://doi.org/10.1051/matecconf/201817503037

10. Pérez-paredes Pascual, Ordoñana-guillamón Carlos (2025) Types of Corpora in Data-Driven Learning. In: *The Palgrave Encyclopedia of Computer-Assisted Language Learning*. Cham: Springer Nature Switzerland. p. 1-10. DOI: 10.1007/978-3-031-51447-0_68-1

11. Murphy Bróna, Riordan Elaine (2016) Corpus types and uses. In: *The Routledge handbook of language learning and technology*. Routledge. p. 388-403. DOI: 10.4324/9781315657899

12. Zhukovska V. V. (2013) Vstup do korpusnoi linhvistyky: navchalnyi posibnyk [Introduction to corpus linguistics: a textbook]. URL: https://eprints.zu.edu.ua/18909/1/korpusna_lingv.pdf [in Ukrainian]

13. Kotsiuk L. M., Kotsiuk Y. A. (2020) Klasifikatsiina paradyhma korpusu tekstiv za osoblyvostiamy yoho dyzainu, struktury ta sposobamy vykorystannia, a takozh sposobom fikatsii ta indeksatsii tekstovykh danykh [Classificational paradigm of a text corpus by its design, structure and use, as well as by the fixation and indexation methods of its text data]. Naukovi zapysky Natsionalnoho universytetu «Ostrovska akademiia». Seriia «Filolohiia», 9 (77). 106-110. URL: http://eprints.oa.edu.ua/id/eprint/9556 [in Ukrainian].

14. Aarts Jan (2022) Corpus analysis. In: *Handbook of Pragmatics*. John Benjamins Publishing Company. p. 1505-1515. DOI: 10.1075/hop.m2.cor2

15. Fang Mingyuan (2023) Forestry english corpus construction and application in foreign language teaching under the background of big data and internet of things. *Journal of Computational Methods in Science and Engineering*, 23.1: 211-221. URL: https://doi.org/10.3233/JCM-226562

16. Lefer Marie-Aude (2021) Parallel corpora. In: *A practical handbook of corpus linguistics*. Cham: Springer International Publishing. p. 257-282. DOI: 10.1007/978-3-030-46216-1_12

17. Lindemann D. Alonso, Arrospide M. (2024) Linking historical corpus data and annotations using Wikibase. Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. – Zagreb: Institute for the Croatian Language. – P. 743–748. URL: https://doi.org/10.5281/zenodo.13932191

18. Coccetta Francesca (2022) Multimodal corpora and concordancing in DDL. *The Routledge handbook of corpora and English language teaching and learning*, 361-376. URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003002901-29/multimodal-corpora-concordancing-ddl-francesca-coccetta

19. Mörth Karlheinz, Dorostkar Niku, Preisinger Alexander (2011) Gleaning micro-corpora from the internet: integrating heterogeneous data into existing corpus infrastructures. *Editores María Luisa Carrió Pastor Miguel Ángel Candel Mora*, 111.

20. Liang Weixin, et al (2022) Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4.8: 669-677. URL: https://doi.org/10.1038/s42256-022-00516-1

21. Chen Haihua, Pieptea Lavinia F., Ding Junhua (2022) Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches. *IEEE Transactions on Reliability*, 71.2: 657-673. DOI: 10.1109/TR.2022.3156126

22. Hou Xiaoqing (2022) Research on translation corpus building with the assistance of ai. In: *2021 International conference on Smart Technologies and Systems for Internet of Things (STS-IOT 2021)*. Atlantis Press. p. 136-140. DOI: 10.2991/ahis.k.220601.027

23. Kovalenko O. S., Serdiuk O. A. (2020) Vykorystannia klasychnykh metodiv mashynnoho navchannia dlia klasifikatsii tekstiv u prohramakh henaratsii avtomatychnoho vidpovidei [Using classical machine learning methods for text classification in automatic response generation programs]. Visnyk Cherkaskoho universytetu. Seriia «Prykladna matematyka. Informatyka», 1. 86-100. URL: https://doi.org/10.31651/2076-5886-2020-1-86-100 [in Ukrainian]

24. Jones John W. (2024) AI, Language, and Education: Bias and Prejudice in Text and Corpus Analysis. In: *Encyclopedia of Educational Innovation*. Springer, Singapore. p. 1-5. URL: https://doi.org/10.1007/978-981-13-2262-4_312-1

25. Ondiba Hesborn (2025) Proactive AI-Driven Cybersecurity for Endangered Language Preservation: Safeguarding the Suba Linguistic Corpus. In: *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*. IEEE. p. 1-5. DOI: 10.1109/ICAIC63015.2025.10848675.

26. Su Zheqian, et al (2020) Influence of "Corpus Data Driven Learning+ Learning Driven Data" Mode on ESP Writing Under the Background of Artificial Intelligence. In: *International Conference on Application of Intelligent Systems in Multi-modal Information Analytics*. Cham: Springer International Publishing. p. 547-552. URL: https://doi.org/10.1007/978-3-030-51431-0_80

27. Gu Feng (2023) Corpus-based critical discourse analysis on AI policy: a comparison between North America and developing countries in East Asia. *Asian J Soc Sci Stud*, 8.3: 14. URL: https://doi.org/10.20849/ajsss.v8i3.1371

28. Mizumoto Atsushi (2023) Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics*, 3.3: 100074. URL: https://doi.org/10.1016/j.acorp.2023.100074

29. Moreno Antonio, Redondo Teófilo (2016) Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 3.6: 57-64. DOI: 10.9781/ijimai.2016.369

Дата першого надходження рукопису до видання: 25.08.2025
Дата прийнятого до друку рукопису після рецензування: 26.09.2025
Дата публікації: 23.10.2025