



Отримано: 13 червня 2026 р.

Прорецензовано: 20 червня 2026 р.

Прийнято до друку: 22 червня 2026 р.

Email: ok.ko176@gmail.com

ORCID ID: <https://orcid.org/0000-0003-1439-7996>

Researcher ID: S-7760-2017

DOI: [http://doi.org/10.25264/2519-2558-2026-30\(98\)-17-21](http://doi.org/10.25264/2519-2558-2026-30(98)-17-21)

Степанечко О. В. Корпус як дзеркало колективного когнітивного досвіду в контексті розвитку мовних моделей (2020–2025). *Наукові записки Національного університету «Острозька академія» : серія «Філологія»*. Острог : Вид-во НаУОА, 2026. Вип. 30(98). С. 17–21.

УДК: 81'322.2:81'23:004.8

Степанечко Оксана Вікторівна,
кандидат педагогічних наук, доцент,
Київський національний університет імені Тараса Шевченка

КОРПУС ЯК ДЗЕРКАЛО КОЛЕКТИВНОГО КОГНІТИВНОГО ДОСВІДУ В КОНТЕКСТІ РОЗВИТКУ МОВНИХ МОДЕЛЕЙ (2020–2025)

У статті представлено комплексний аналіз розвитку корпусної лінгвістики в Україні крізь призму лінгвокогнітивного виміру в контексті створення великих мовних моделей (LM) у період 2020–2025 років. У дослідженні аналізується еволюція українських корпусних ресурсів, методів лінгвістичної анотації й інструментів обробки природної мови, що формують основу для навчання сучасних мовних моделей. Особлива увага звертається на динамічну трансформацію корпусних даних та появу вітчизняних архітектур штучного інтелекту, зокрема MamayLM і Lapa LLM, що сприяють формуванню цифрової мовної картини світу українського суспільства та відображають специфіку національного комунікативного простору. Авторка обґрунтовує концепцію лінгвістичного корпусу як об'єктивного «дзеркала» колективного когнітивного досвіду, що зосереджує ментальні структури, культурні коди та ціннісні орієнтири нації в умовах глобальних трансформацій і цифровізації. У дослідженні визначено ключові лінгвістичні виклики української мови для систем обробки природної мови, а саме: висока флективність, морфологічна варіативність і відносно вільний порядок слів, що постають не лише як технічні труднощі для NLP-моделей, а й як прояв специфічних когнітивних механізмів організації мовлення. Проаналізовано сучасні методології корпусної розмітки, включно з багаторівневою анотацією граматичних, семантичних і емоційних параметрів тексту, що дозволяє мовним моделям точніше інтерпретувати контекст і прагматику українського мовлення. Доведено, що розвиток національних корпусів і великих мовних моделей є важливим чинником забезпечення когнітивної безпеки та цифрового суверенітету України, оскільки сприяє репрезентації унікального культурного й мовного досвіду в глобальному інформаційному просторі та формує підґрунтя для подальшого розвитку українських технологій штучного інтелекту.

Ключові слова: корпусна лінгвістика, лінгвокогнітивний вимір, мовні моделі (LLM), когнітивні структури, українська мова, NLP, анотаційні схеми.

Oksana Stepanechko,
Ph.D. in Pedagogy, Associate Professor,
Taras Shevchenko National University of Kyiv

CORPUS AS A MIRROR OF COLLECTIVE COGNITIVE EXPERIENCE IN THE CONTEXT OF LANGUAGE MODELS DEVELOPMENT (2020–2025)

In the context of developing large language models (LLMs) between 2020 and 2025, the article provides a thorough examination of the evolution of corpus linguistics in Ukraine through the lens of the linguo-cognitive dimension. The research investigates the development of natural language processing tools, linguistic annotation techniques, and Ukrainian corpus resources that serve as the foundation for training contemporary language models. The emergence of domestic artificial intelligence architectures, particularly MamayLM and Lapa LLM, and the dynamic transformation of corpus data are highlighted as ways that they reflect the national communicative space and contribute to the development of a digital linguistic worldview of Ukrainian society. The idea that the linguistic corpus is an objective “mirror” of collective cognitive experience that accumulates mental structures is grounded by the author. The article identifies key linguistic challenges of the Ukrainian language for natural language processing systems, including high inflectional complexity, morphological variability, and flexible syntax. These features show up as manifestations of particular cognitive mechanisms of speech organization rather than just as technical challenges for NLP models. In order to help language models better understand the context and pragmatics of Ukrainian speech, the study examines contemporary corpus labelling techniques, such as multi-level annotation of grammatical, semantic, and emotional text parameters. It is argued that the creation of national corpora and large language models is an essential factor in ensuring the cognitive security and digital sovereignty of Ukraine, as it facilitates the representation of unique cultural and linguistic experiences in the global information space and forms the foundation for the further development of Ukrainian artificial intelligence technologies.

Keywords: corpus linguistics, linguo-cognitive dimension, Large Language Models (LLMs), cognitive structures, Ukrainian language, NLP, annotation schemes.

Беззаперечно, що весь світ знаходиться в епіцентрі технологічного й техногенного прориву й нестримного розвитку штучного інтелекту (ШІ), який уже став невід'ємною складовою нашого життя, і, починаючи з 2022 року, ШІ почав самонавчатися, саморозвиватися й самовдосконалюватися. Усі ці процеси є не лише технічним, а й когнітивним викликом. Незважаючи на наше суб'єктивне ставлення до змін, що супроводжуються технологічним проривом, раціональним рішенням є опанування цих інструментів і уміле користування ними з метою покращення умов життя й розвитку, як особистісного, так і державного. Водночас, ШІ починає стимулювати людські процеси самонавчання, що потребують глибшого розуміння того, як мова кодує наш досвід.

Історичні події й війна з російською федерацією спричинили докорінні зміни в колективній свідомості українців, посиливши використання рідної мови як головного інструменту ідентифікації й пізнання. Водночас цифровізація країни державною мовою створює умови для фіксації унікальної української мовної картини світу, що характеризується багатоморфологією, флективністю та гнучким синтаксисом. Усі ці чинники сприяють розробці власних мовних моделей (LM –



language models) в Україні і є кроком до когнітивної та технологічної незалежності. Це говорить про те, що Україна активно розвиває власний ШІ-суверенітет, створюючи технологічну свободу та забезпечуючи захист національних даних. І як восьма країна світу, що активно розробляє LM для національної мови, держава потребує науково обґрунтованих стратегій оптимізації ресурсів.

Завдання дослідження:

1. Систематизація існуючих корпусних ресурсів української мови, створених чи розширених задля потреб навчання мовних моделей (2020–2025).

1. Визначення специфічних лінгвістичних викликів української мови (морфологічна варіативність, синтаксична гнучкість, діалектне розмаїття), що впливають на якість LM.

2. Аналіз нових лінгвістичних методологій та анотаційних схем, що виникли унаслідок потреб LM (sentiment analysis, NER) для української мови.

3. Ідентифікація прогалин у корпусному покритті української мови й пропозиції напрямів майбутніх досліджень.

4. Обґрунтування ролі відкритих корпусних ресурсів у забезпеченні лінгвістичного суверенітету в епоху ШІ.

Наукова новизна дослідження: уперше досліджено двосторонній зв'язок між розвитком LM та корпусною лінгвістикою саме для української мови в період з 2020 до 2025 рр. Виявлено специфічні патерни «корпусного голоду» для рідної мови та шляхи його подолання.

Розвиток комп'ютерних технологій спричинив появу нових можливостей для відбору й аналізу функціональних характеристик мовних одиниць у різних типах текстів. Корпусна лінгвістика стає одним із пріоритетних інструментів сучасних лінгвістичних досліджень і зосереджується на аналізі природної мови в умовах реального функціонування з використанням комп'ютерних технологій на основі великих за обсягом, ретельно відібраних і впорядкованих текстових корпусів. Як стверджує Н. Таценко (Таценко, 2020): «Попит на корпусні дані збігся з появою відповідних технічних можливостей». Корпусна лінгвістика сьогодні переорієнтовує традиційні підходи, дозволяючи вивчати мову як об'єкт, що відображає ментальний досвід через реальні тексти, а «результати аналізу даних корпусу сприяють переорієнтації низки лінгвістичних теорій» (MacEnery, Hardie, 2012: 1). На думку О. Демської (Демська, 2016: 33) «аби електронні текстові зібрання можна було вважати корпусом, важливо дотримуватися основних критеріїв: репрезентативність, стандартність, вичерпність». На сьогодні зафіксовано понад 600 комп'ютерних створюваних корпусів.

Емпіричний аналіз даних забезпечує об'єктивне пізнання мовних явищ, усуваючи дослідницькі упередження щодо того, як функціонує мислення. Корпусні студії орієнтовані на вивчення природної мови в умовах її дійсного функціонування із застосуванням комп'ютерних технологій та спираються на великі за обсягом, систематизовані й ретельно відібрані текстові корпуси. А. Алексєєв (Алексєєв 2023: 6) доводить, що «ґрунтуючись переважним чином на емпіричному аналізі лінгвістичних даних, корпусний підхід дає змогу мінімізувати вплив дослідницької суб'єктивності та забезпечує більш об'єктивне й верифіковане пізнання мовних явищ».

За визначенням В. Жуковської (Жуковська 2013: 9) «корпусна лінгвістика як галузь прикладного мовознавства визначає загальні принципи побудови, обробки та використання лінгвістичних корпусів із застосуванням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження й аналізу». Застосування комп'ютерів дає змогу миттєво обробити величезний обсяг мовного матеріалу й відібрати всі можливі в конкретному корпусі приклади вживання необхідних для аналізу одиниць (Жуковська 2013: 13).

«Корпусні дослідження базуються на великих, упорядкованих і анотованих текстових масивах, що відповідають критеріям репрезентативності й об'єктивності лінгвістичної інформації» (Нахлік & Ільницька, 2023). Якість корпусу оцінюється за кількістю текстів, поширеністю жанрів і стилів, частотністю мовних одиниць, рівнем анотації (морфологічної, синтаксичної, семантичної) і наявністю метаданих для аналізу за часом, жанром чи соціолінгвістичними ознаками. Репрезентативний корпус має відображати реальне функціонування мови та містити достатню кількість контекстів для кожної мовної одиниці, що забезпечує емпіричну повноту та достовірність висновків.

Сучасні підходи підкреслюють необхідність стандартизованої розмітки й автоматизованих інструментів обробки, що підвищує порівнюваність і відтворюваність результатів. Поєднання комп'ютерної обробки з лінгвістичною експертизою дозволяє створювати корпуси із багаторівневими анотаціями та аналізувати матеріал на різних лінгвістичних рівнях. У морфологічно складних мовах, таких як українська, методологічно важливо забезпечити «лінгвістичну повноту», що включає не лише кількість токенів і типів, а й якісне представлення морфологічних форм, контекстів і синтаксичних конструкцій.

Особливості морфології, синтаксису й діалектної варіативності української мови створюють додаткові виклики для розробки мовних моделей, тому сучасні методології анотації адаптуються до цих характеристик для підвищення ефективності NLP-систем.

«Україна прагне увійти до трійки світових лідерів з упровадження штучного інтелекту, насамперед шляхом розробки власної великої мовної моделі (LLM). LLM є інструментом, що розуміє національний контекст, діалекти й терміни української мови (EPAM, 2025).

Основою для навчання української LLM є наявні корпуси текстів. «Корпус української мови» на порталі MOVA.info вважається найбільшим ресурсом і містить понад 100 мільйонів слововживань. Він розроблений у лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка» (EPAM, 2025).

«Реалізація проєктів «Корпуси української мови» набуває стратегічного значення державного, наукового та суспільного рівня» (Демська-Кульчицька, 2008), адже йдеться не лише про створення інструменту для лінгвістичних досліджень, а й про фіксацію, збереження й репрезентацію української мовної картини світу в цифровому просторі.

У сучасних умовах ці проєкти усе більше розглядаються як елемент цифрового суверенітету та культурної безпеки, особливо з огляду на тривалі утиски нашої мови в інформаційних технологіях у минулі десятиліття. Наявність повноцінного національного корпусу забезпечує проявлення української мови в глобальних наукових і технологічних процесах, зокрема у сфері штучного інтелекту, автоматичної обробки природної мови, цифрової освіти та гуманітарних досліджень. «Національна LLM стає потужним технологічним проривом для України, відкриваючи широкі можливості для розробки AI-рішень» (EPAM, 2025).

Період 2020–2025 років характеризується інтенсивним розвитком великих мовних моделей (Large Language Models, LLM) для української мови, що створює нові виклики та можливості для корпусної лінгвістики. Кількість відкритих моделей, що працюють з українською мовою, зросла на 122% за останній рік – це найвищий показник серед усіх мов на платформі Hugging Face (Чаплинська, 2025).

На сучасному етапі спостерігається поступовий розвиток корпусних ініціатив для української мови. Наймасштабнішим проектом вважається ГРАК (Генеральний регіонально анотований корпус української мови (GRAC)).

ГРАК (ГРАК 2017-2025) – «це велика репрезентативна добірка текстів, написаних українською мовою, доступна через складний інтерфейс», що налічує 1,9 мільярда токенів. «У ГРАКу подано реальне використання сучасної української мови за широкий проміжок часу з охопленням розлогого діапазону географічних, стилістичних та інших варіацій. Корпус містить понад 150 тисяч текстів, що їх написали близько 35 тисяч авторів за період від 1816 року до наших днів» (ГРАК, 2017-2025). Корпус призначений для здійснення лінгвістичних досліджень із граматики, лексики, історії української літературної мови, а також для використання під час укладання словників, підручників і граматик.

Проект «Малюк» є найбільшим відкритим українським корпусом, створеним спеціально для навчання мовних моделей (Hugging Face). Однак, як оцінюють експерти, обсяг цього корпусу необхідно збільшити щонайменше у сім разів, аби досягти рівня, придатного для повноцінного навчання сучасних LLM.

Проблема обсягу корпусів є критичною для української мови. Якщо порівняти з англійськими корпусами, де доступні терабайти текстових даних, українські ресурси залишаються обмеженими. Це пов'язано не лише з технічними, а й з юридичними питаннями авторського права та доступу до текстів.

Для навчання передових українських LLM використовуються й інші корпуси обсягом 75 мільярдів токенів українською й англійською мовами.

Та попри величезну роботу із розвитку україномовних моделей, складність навчання LM для української мови зумовлена специфікою того, як носії мови структурують інформацію, адже окрім того, що нашій мові притаманні символічність і полісемантичність (Пацай, 2025:173), ще й:

- морфологічна варіативність проявляється у високій гнучкості: до 7 відмінків, 3 роди, 2 числа й аспекти дієслів та дієвідмінювання створюють численні словоформи для кожної лексеми, що свідчить про складну систему категоризації понять у свідомості, де кожна зміна форми слова несе своє когнітивне навантаження. Водночас це ускладнює токенізацію та лематизацію в LM, призводячи до помилок у розпізнаванні значень через омонімію форм. Якщо в англійському тексті один токен відповідає 3–4 символам, то в українській мові через багату морфологію одне слово розбивається на 2–4 токени. У свою чергу, морфологічна варіативність впливає на розмір словника моделі – потрібно враховувати всі можливі форми слів; якість токенізації через необхідність розпізнавання морфологічних меж; узгодження слів у реченні – складніші правила, ніж в англійській мові.

- вільний порядок слів у реченні (SVO (Subject-Verb-Object), SOV, OVS) показує, що в українській когнітивній моделі роль об'єкта чи суб'єкта визначається внутрішньою формою слова, а не чіткою структурою, що створює додаткові виклики для навчання мовних моделей й вимагає глибшого розуміння контексту. Така синтаксична гнучкість впливає на модель шляхом збільшення варіативності можливих конструкцій; ускладнення визначення синтаксичних ролей слів; потреби більших обсягів даних для навчання.

- наявність понад 20 діалектів (північне, південно-західне, східне) із фонетичними, лексичними й морфологічними варіаціями відображає багаторівневий досвід сприйняття світу, який моделі мають навчитися інтерпретувати без упереджень до літературної норми. Для мовних моделей це означає необхідність обробляти варіанти мови з різних регіонів України та діаспори. ГРАК враховує цю особливість через регіональну анотацію текстів.

Сучасні дослідження обробки природної мови для української мови спираються на розвинену систему корпусних ресурсів, анотаційних стандартів і бенчмарків. «Забезпечення правильної та якісної роботи NLP є критично важливим завданням» (Пацай, 2025: 173). У межах стандарту Universal Dependencies (Universal Dependencies, 2024) застосовується універсальна система тегування UPOS, доповнена морфологічними ознаками, що враховують флексивну специфіку української мови. На основі цих принципів створюються корпуси та набори даних для задач розпізнавання іменованих сутностей (NER) і аналізу тональності тексту (sentiment analysis), де використовуються багаторівневі схеми анотації, що враховують контекстуальні, жанрові та морфологічні характеристики мовлення, зокрема вплив граматичних категорій.

Значний внесок у розвиток українських NLP-ресурсів зробили масштабні корпуси, зокрема UberText, «зібрані й упорядковані на основі текстів українських періодичних видань» (Корпуси текстів, 2024), а також набори даних для розпізнавання сутностей NER-UK, що використовуються для навчання та оцінювання мовних моделей (Chaplynskyi, 2023; Chaplynskyi & Romanushyn, 2024). Паралельно розробляються «спеціалізовані бенчмарки для тестування великих мовних моделей, зокрема ZNO-Eval, що базується на завданнях Зовнішнього незалежного оцінювання й дозволяє оцінювати здатність моделей працювати з академічними текстами, демонструвати розуміння прочитаного, логіку мовлення та граматичну точність у межах українського освітнього контексту» (Sygomiatiukov, Ruvinskaya & Troynina, 2024).

Упродовж 2020–2025 рр. спостерігається стрімкий розвиток великих мовних моделей (LLM), адаптованих до української мови. Значна частина цих систем створюється на основі відкритих архітектур та донавається на великих корпусах україномовних текстів. Такі моделі орієнтовані на покращення якості обробки української мови, включаючи аналіз тексту, переклад, створення контенту й підтримку освітніх і державних цифрових сервісів.

MamaLM – це одна з найсучасніших великих мовних моделей, адаптованих для української мови, розроблена дослідниками інституту INSAIT (Болгарія) у співпраці з ETH Zurich (MamaLM-Gemma, 2024). Її основними характеристиками є: 1) двомовна архітектура (українська та англійська мови); 2) адаптація до граматичних та культурних особливостей української мови; 3) висока продуктивність у тестах, зокрема на завданнях Зовнішнього незалежного оцінювання, де модель дозволяє перевірити не просто знання слів, а й її здатність до логіки та смислових зв'язків; 4) відкритий доступ до моделі через платформу Hugging Face (Dev.ua, 2024).

Завдяки відносно компактному розміру (9 млрд параметрів) модель може запускатися навіть на одному графічному процесорі, що робить її придатною для використання в освітніх і державних установах (Dev.ua, 2024).



Лара LLM (версія 0.1.2, 2024) є однією з перших відкритих мовних моделей, спеціально розроблених для української мови. Вона побудована на базовій архітектурі Gemma 3 й оптимізована для задач машинного перекладу, генерації тексту і мовного аналізу. Модель демонструє високі результати у задачах перекладу, перевершуючи низку інших відкритих моделей аналогічного класу, а також в обробці зображень українською. Окрім того, дана модель успішно проходить тести на визначення пропаганди й дезінформації завдяки спеціальним фільтраційним методам. Вихідний код та ваги моделі доступні на платформах Hugging Face та GitHub, що забезпечує її використання в дослідницьких та освітніх проєктах (Лара LLM 2025).

Окремим напрямом розвитку є створення національної мовної моделі за підтримки Міністерства цифрової трансформації України. Проєкт спрямований на формування масштабної українськомовної LLM для використання в державному секторі, освіті, обороні та бізнесі.

Основні характеристики проєкту: 1) використання відкритої архітектури, зокрема моделей сімейства Gemma; 2) попереднє навчання моделі на українських корпусах даних; 3) фінансування за участю приватних бізнес-партнерів без залучення державного бюджету; 3) забезпечення безоплатного доступу для державних установ, університетів, шкіл і науковців (Міністерство цифрової трансформації України, 2025).

Очікується, що ця модель стане ключовою інфраструктурою для розвитку українських ІІІ-сервісів і сприятиме інтеграції штучного інтелекту в державне управління та освітні системи.

Попри успіхи, існує проблема «корпусного голоду», оскільки більшість корпусів побудовані на основі писемних текстів і майже не відображають усне мовлення. Для досягнення «лінгвістичної повноти», яка б повноцінно відображала все різноманіття людського досвіду, необхідно: збільшити обсяги корпусів у 5–7 разів для охоплення академічних, художніх та технічних сфер; подолати брак оцифрованого усного мовлення, що є первинним джерелом вивчення когнітивних процесів «тут і зараз»; стандартизувати метадані для кращої порівнюваності результатів когнітивних розвідок; інтегрувати українські корпуси в міжнародні проєкти.

На основі викладеного вище, можемо дійти таких висновків: розвиток корпусної лінгвістики в Україні (2020–2025) трансформувався з технічної задачі у фундаментальне дослідження лінгвокогнітивного виміру. Створення національних LLM є спробою інтегрувати український когнітивний досвід у глобальний цифровий простір, забезпечуючи лінгвістичний суверенітет та збереження ідентичності в епоху ІІІ.

Україна демонструє найвищі темпи зростання кількості мовних моделей серед усіх мов (122% за останній рік), що свідчить про активний розвиток галузі. Однак для подальшого розвитку необхідно вирішити питання збільшення обсягів корпусів, юридичного регулювання використання текстів та створення спеціалізованих анотованих ресурсів.

Таким чином, сучасний стан корпусів української мови можна охарактеризувати як етап активного формування, що поєднує значні наукові здобутки з потребою в системній державній підтримці, стандартизації та довготривалому фінансуванні. Подальший розвиток національного корпусу української мови є не лише лінгвістичним завданням, а й важливим кроком у напрямі утвердження української мови як повноцінного суб'єкта глобального цифрового простору.

Перспективи майбутніх розвідок у царині корпусної лінгвістики включають розширення корпусної бази, розробку автоматизованих інструментів анотації, створення спеціалізованих корпусів та активну міжнародну співпрацю в галузі NLP для української мови.

Література:

1. Алексєєв А. Взаємодія знань корпусної лінгвістики та інформаційних технологій у навчальному процесі. *Сучасна прикладна лінгвістика: зб. наук. пр.* Дніпро: ВНІЗ «ДГУ», 2023. Ч. II. С. 5–14.
2. В Україні представили велику мовну модель Lapa LLM: що відомо. *MediaSapiens*. 2025. URL: <https://ms.detector.media/internet/post/38562/2025-10-27-v-ukraini-predstavlyu-veliku-movnu-model-lapa-llm-shcho-vidomo/>
3. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко, Т. Ніколаєнко, А. Лукашевський та ін. – Київ, Львів, Єна, 2017–2025. <https://uacorporus.org>
4. Голошук С. Корпусна лінгвістика: сучасний стан та перспективи досліджень. *Current Issues of Linguistics and Translation Studies*. 2021. № 22. С. 33–36. DOI: <https://doi.org/10.31891/2415-7929-2021-22-7>.
5. Демська О. М. Українська корпусна лінгвістика: сьогодення і перспективи. *Дивослово*. 2016. № 6. С. 32–34.
6. Демська-Кульчицька О. Український національний корпус [Електронний ресурс] / Оріся Демська-Кульчицька // Проєкт Растко : бібліотека сербської культури на інтернеті. – Режим доступу: <https://www.rastko.rs/rastko/delo/12244>
7. Жуковська В. В. *Вступ до корпусної лінгвістики: навч. посібник*. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с.
8. Корпус української мови [Електронний ресурс] : [вебсайт]. – Режим доступу: <http://www.mova.info/corpus.aspx>
9. Корпуси тексту // Лабораторія комп'ютерної лінгвістики. мова.інфо : [інформ. портал]. 2024. URL: <https://lang.org.ua/uk/corpora/>
10. Міністерство цифрової трансформації України. (2025, December 1). *Українську LLM напрують на моделі Gemma від Google*. <https://thedigital.gov.ua/news/technologies/ukrayinsku-llm-natrenuiut-na-modeli-gemma-vid-google>
11. Нахлік О., Ільницька Ю. Корпусно-базований підхід дослідження дієслів семантичної категорії «мовлення/communication» в українській та англійській мовах. *Молодий вчений*. 2023. № 10 (122). С. 60–66. DOI: <https://doi.org/10.32839/2304-5809/2023-10-122-13>.
12. Пацай Б. Обробка природної мови українською: виклики та перспективи використання штучного інтелекту в освіті [Текст] / Б. Пацай, І. Нечипорук, А. Ковтун // *Цифрова економіка та економічна безпека : науково-практичний журнал* / [гол. ред. О. Ю. Кудріна, редкол.: В. В. Божкова, В. І. Борщ, Н. М. Вдовенко та ін.]. – 2025. – № 1 (16). – С. 172–179. – DOI: <https://doi.org/10.32782/dees.16-26>
13. Розробляємо українську велику мовну модель (LLM). *Міністерство цифрової трансформації України*. 2025. URL: <https://thedigital.gov.ua>
14. Руденко О. В. Проблеми створення та використання національних корпусів мов. *Мовознавство*. 2012. № 4. С. 45–53.
15. Створення національної LLM: виклики, переваги та шлях до мовної незалежності. *EPAM. Blog*. 10.06.2025. URL: <https://careers.epam.ua/blog/creating-a-national-llm>
16. Таценко Н. В. Методи корпусної лінгвістики в підготовці фахівців філологів. *Scientific and Pedagogical Internship "Organization of educational process in the field of philological sciences in Ukraine and EU countries": Internship proceedings (August 24 – October 2)*. Venice: Izdavnica "Baltija Publishing", 2020. С. 177–181.
17. Українські розробники випустили відкриту мовну модель «Мамай». *dev.ua*. 2024. URL: <https://dev.ua/news/mama-1745405680>

18. Чаплинська Ю. Частка української мови в навчаннях ШІ-моделей зростає – дослідження. *DOU*. 2025. URL: <https://dou.ua/lenta/news/ukrainian-language-is-growing-among-ai-models/>
19. Широков В. А., Шевченко І. С., Рабулець О. В. *Корпусна лінгвістика: монографія*. Київ: Довіра, 2005. 472 с.
20. Chaplynskyi D. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. 2023. p. 1-10.
21. Chaplynskyi D., Romanyshyn M. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian. *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, 23-29.
22. Krylovych O., Shyrokov V. General Regionally Annotated Corpus of Ukrainian (GRAC). *Proceedings of the International Conference on Corpus Linguistics*. 2018. P. 112–120.
23. MamayLM-Gemma-2-9B-IT-v0.1. *Hugging Face*. 2024. URL: <https://huggingface.co/INSAIT-Institute/MamayLM-Gemma-2-9B-IT-v0.1>
24. McEnery T., Wilson A. *Corpus linguistics: an introduction*. 2nd ed. Edinburgh: Edinburgh University Press, 2001. 235 p.
25. Syromiatnikov R., Ruvinskaya V., Troynina A. ZNO-Eval: Benchmarking reasoning capabilities of large language models in Ukrainian. *X International conference "Informatics. Culture. Technology."* 2024. p. 185-191. <https://doi.org/10.48550/arXiv.2501.06715>
26. Ukrainian UD treebanks // Universal Dependencies. 2024. URL: <https://universaldependencies.org/uk/index.html>

References:

1. Aliekseiev, A. (2023). Vzaemodiia znan korpusnoi linhvistyky ta informatsiinykh tekhnolohii u navchalnomu protsesi [Interaction of corpus linguistics knowledge and information technologies in the educational process]. *Suchasna prykladna linhvistyka*, 2, 5–14.
2. Chaplynska, Y. (2025, January 23). *Chastka ukrainskoi movy v navchanniakh ShI-modelei zrostaie – doslidzhennia* [The share of the Ukrainian language in AI model training is growing – research]. *DOU*. <https://dou.ua/lenta/news/ukrainian-language-is-growing-among-ai-models/>
3. Chaplynskyi, D. (2023). Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, 1–10.
4. Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian. *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, 23–29.
5. Demska, O. M. (2016). Ukrainska korpusna linhvistyka: siohodennia i perspektyvy [Ukrainian corpus linguistics: Present and prospects]. *Dyvoslovo*, (6), 32–34.
6. Demska-Kulchytska, O. (n.d.). *Ukrainskyi natsionalnyi korpus* [Ukrainian National Corpus]. Rastko Project. <https://www.rastko.rs/rastko/delo/12244.dev.ua>. (2024, May 15).
7. EPAM Ukraine. (2025, June 10). *Stvorennia natsionalnoi LLM: vyklyky, perevahy ta shliakh do movnoi nezalezhnosti* [Creating a national LLM: Challenges, advantages and the path to language independence]. EPAM Blog. <https://careers.epam.ua/blog/creating-a-national-llm>
8. Holoshchuk, S. (2021). Korpusna linhvistyka: suchasnyi stan ta perspektyvy doslidzen [Corpus linguistics: Current state and research prospects]. *Current Issues of Linguistics and Translation Studies*, (22), 33–36. <https://doi.org/10.31891/2415-7929-2021-22-7>
9. Hugging Face. (2024). *MamayLM-Gemma-2-9B-IT-v0.1*. <https://huggingface.co/INSAIT-Institute/MamayLM-Gemma-2-9B-IT-v0.1>
10. Krylovych, O., & Shyrokov, V. (2018). General Regionally Annotated Corpus of Ukrainian (GRAC). *Proceedings of the International Conference on Corpus Linguistics*, 112–120.
11. Laboratory of Computational Linguistics. (n.d.). *Corpora of texts*. Mova.info. <https://lang.org.ua/uk/corpora/>
12. McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.
13. MediaSapiens. (2025, October 27). *V Ukraini predstavlyly velyku movnu model Lapa LLM: shcho vidomo* [Large language model Lapa LLM presented in Ukraine: What is known]. <https://ms.detector.media/internet/post/38562/2025-10-27-v-ukraini-predstavlyly-velyku-movnu-model-lapa-llm-shcho-vidomo/>
14. Ministry of Digital Transformation of Ukraine. (2025). *Rozrobluiemo ukrainsku velyku movnu model (LLM)* [Developing the Ukrainian Large Language Model (LLM)]. <https://thedigital.gov.ua>
15. Ministry of Digital Transformation of Ukraine. (2025, December 1). *Ukrainsku LLM natrenuiut na modeli Gemma vid Google* [Ukrainian LLM will be trained on Google's Gemma model]. <https://thedigital.gov.ua/news/technologies/ukrayinsku-llm-natrenuiut-na-modeli-gemma-vid-google>
16. Mova.info. (n.d.). *Korpus ukrainskoi movy* [Corpus of the Ukrainian language]. <http://www.mova.info/corpus.aspx>
17. Nakhlik, O., & Ilnytska, Yu. (2023). Korpusno-bazovanyi pidkhid doslidzhennia diiesliv semantichnoi katehorii "movlennia/communication" v ukrainskii ta anhliiskii movakh [Corpus-based approach to the study of verbs of the semantic category "speech/communication" in Ukrainian and English]. *Molodyi vchenyi*, 10(122), 60–66. <https://doi.org/10.32839/2304-5809/2023-10-122-13>
18. Patsai, B., Nechyporuk, I., & Kovtun, A. (2025). Obrobka pryrodnoi movy ukrainskoii: vyklyky ta perspektyvy vykorystannia shtuchnoho intelektu v osviti [Natural language processing in Ukrainian: Challenges and prospects of using AI in education]. *Tsyfrova ekonomika ta ekonomichna bezpeka*, 1(16), 172–179. <https://doi.org/10.32782/dees.16-26>
19. Rudenko, O. V. (2012). Problemy stvorennia ta vykorystannia natsionalnykh korpusiv mov [Problems of creation and use of national language corpora]. *Movoznavstvo*, (4), 45–53.
20. Shviedova, M., von Waldenfels, R., Yaryhin, S., Rysyn, A., Starko, V., Nikolaienko, T., Lukashevskiy, A., et al. (2017–2025).
21. *Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK)* [General Regionally Annotated Corpus of Ukrainian (GRAC)]. <https://uacorpus.org>
22. Shyrokov, V. A., Shevchenko, I. S., & Rabulets, O. V. (2005). *Korpusna linhvistyka: Monohrafiia* [Corpus linguistics: Monograph]. Dovira.
23. Syromiatnikov, R., Ruvinskaya, V., & Troynina, A. (2024). ZNO-Eval: Benchmarking reasoning capabilities of large language models in Ukrainian. *X International conference "Informatics. Culture. Technology."*, 185–191. <https://doi.org/10.48550/arXiv.2501.06715>
24. Tatsenko, N. V. (2020). Metody korpusnoi linhvistyky v pidhotovtsi fakhivtsiv filolohiv [Methods of corpus linguistics in the training of philologists]. *Scientific and Pedagogical Internship "Organization of educational process in the field of philological sciences in Ukraine and EU countries"*, 177–181.
25. Ukrainski rozrobnyky vypustyly vidkrytu movnu model "Mamai" [Ukrainian developers released the open language model "Mamai"]. <https://dev.ua/news/mama-1745405680> dev.ua. (2024, June 10).
26. Universal Dependencies. (2024). *Ukrainian UD treebanks*. <https://universaldependencies.org/uk/index.html>
27. Zhukovska, V. V. (2013). *Vstup do korpusnoi linhvistyky: Navch. posibnyk* [Introduction to corpus linguistics: Tutorial]. ZhDU im. I. Franka.