УДК 004.85+81’322

ГІБРИДНА ЛІНГВІСТИЧНО ОРІЄНТОВАНА МОДЕЛЬ МАШИННОГО ПЕРЕКЛАДУ АНГЛО-УКРАЇНСЬКИХ ТЕКСТІВ

І. М. Лях, М. П. Кляп, А. Ю. Ціпіньо, А. М. Роман

Ужгородський національний університет,
факультет інформаційних технологій,
пл. Народна, 3, Ужгород, 88000, Україна

У статті розглянуто еволюційний розвиток машинного перекладу – від правило-орієнтованих систем (RBMT) і статистичних методів (SMT) до нейронних архітектур нового покоління (NMT). Особливу увагу приділено порівняльному аналізу морфологічних, синтаксичних і семантичних відмінностей між англійською та українською мовами, які зумовлюють обмеженість точності автоматизованих систем. Серед ключових розбіжностей вказано флексійну варіативність, відмінності у порядку слів, граматичному роді, числі, виді та категорії узгодження, що створюють лінгвістичні бар’єри для машинного відтворення природного українського синтаксису.

Для вирішення окреслених проблем запропоновано гібридну лінгвістично орієнтовану модель перекладу, яка інтегрує механізми морфосинтаксичного аналізу з нейронною мережею. У статті детально описано поетапну структуру системи: блок попередньої морфосинтаксичної нормалізації, модуль екстракції лінгвістичних ознак, нейронне ядро перекладу, український морфогенератор і підсистему постредагування. Задля об’єктивної оцінки ефективності запропонованої архітектури передбачено використання міжнародних метрик оцінки якості – BLEU (Bilingual Evaluation Understudy) і TER (Translation Edit Rate), що дозволяють визначити точність і природність отриманого перекладу порівняно з еталонним людським.

Автори обґрунтовують доцільність залучення граматичних і контекстних правил як складової нейронних моделей для перекладу флективних мов, зокрема української. Такий підхід спрямований на вдосконалення точності, стилістичної природності та контекстуальної адекватності комп’ютерних перекладацьких систем. Результати дослідження визначають перспективу подальших розробок у напрямі створення великих паралельних англо-українських корпусів, апробації запропонованої моделі в експериментальному середовищі та розширення підходу до багатомовних перекладацьких платформ. У практичному вимірі це сприятиме формуванню якісних україномовних перекладацьких інструментів, інтегрованих у сучасні системи обробки природної мови.

Ключові слова: *автоматизований переклад; машинний переклад; нейронна мережа; морфосинтаксичний аналіз; BLEU; TER.*

Постановка проблеми. У сучасних умовах глобалізації та цифровізації обмін інформацією між мовними спільнотами значною мірою залежить від ефективності

перекладу. Машинний переклад (МП) та автоматизовані системи підтримки перекладу стають важливим інструментом для забезпечення швидкого й точного обміну знаннями, особливо в науковій, освітній, економічній та медійній сферах. У науковій літературі широко досліджуються різні підходи до машинного перекладу – від правило-орієнтованих систем перекладу (Rule-based machine translation, RBMT) та статистичних моделей перекладу (Statistical machine translation, SMT) до сучасних нейронних моделей перекладу (Neural machine translation, NMT) та гібридних систем. Крім того, увага приділяється проблемам адекватності перекладу, особливостям англо-українського перекладу, застосуванню сучасних архітектур, таких як: Transformer, BERT, MarianMT, а також питанням оцінювання якості перекладу за допомогою автоматичних та експертних метрик. Аналіз існуючих досліджень дозволяє виділити ключові теоретичні основи, практичні стратегії та напрямки подальшого розвитку перекладознавства та машинного перекладу, зокрема для мов із обмеженими ресурсами, до яких належить українська.

Аналіз останніх досліджень та публікацій. Автори досліджують можливості великих мовних моделей, зокрема ChatGPT, у сфері машинного перекладу, зосереджуючись на лексичних, семантичних та синтаксичних особливостях перекладу академічних текстів з англійської на українську. Результати показують, що хоча модель демонструє природну і точну генерацію тексту, вона все ще робить помилки з комплексними мовними структурами, що обмежує її здатність повністю замінити професійного перекладача. Стаття підкреслює цінність великих мовних моделей як інструменту для аналізу текстових патернів і покращення машинного перекладу, водночас вказуючи на існуючі обмеження [1]. Литвин та інші дослідники розробили модель машинного навчання для виправлення граматичних помилок в українських текстах, застосовуючи передтреновані трансформери BERT та mT5. Вони показали, що система добре працює з простими реченнями, проте для повноцінної корекції необхідне поєднання словників і правил залежностей. Дослідження демонструє потенціал трансформерів у виправленні граматичних та стилістичних помилок, одночасно підкреслюючи обмеження при обробці складних конструкцій [2]. Mandziy та інші дослідники досліджують застосування паралельного корпусу англійсько-українських ІТ-текстів у перекладознавстві. Вони описують створення корпусу за допомогою Sketch Engine та розробку макросу для MS Word, який полегшує пошук і правильний переклад ІТ-термінів. Стаття демонструє теоретичну цінність у систематизації перекладу термінології та практичну користь для лінгвістів, перекладачів і ІТ-фахівців, забезпечуючи більш точний і ефективний переклад спеціалізованих текстів [3].

Vysotska V. розробила модель комп'ютерної лінгвістичної системи (Computer linguistic system, CLS) для обробки українських текстів, яка враховує всі рівні аналізу – від графемного до прагматичного. Дослідження підкреслює адаптацію NLP-процесів до складної морфології української мови та показує приклади вирішення типових задач, таких як ідентифікація «вірусних» заголовків і виправлення граматичних та стилістичних помилок. Стаття демонструє комплексний підхід до моделювання систем обробки української мови та виділяє ключові структурні

елементи для подальшого розвитку CLS [4]. У статті Che C. розглянуто застосування алгоритму диференціальної еволюції для побудови та моделювання інтерактивної моделі навчання англійського перекладу з метою підвищення ефективності викладання та розвитку інтерактивності освітнього процесу. Автори відзначають, що в умовах онлайн-навчання традиційний контроль за успішністю студентів замінюється аналізом їхньої навчальної активності – логінів, переглядів, виконання завдань, комунікації та тестів. Алгоритм диференціальної еволюції ефективно аналізує та оптимізує структуру навчальних даних, оцінює вплив налаштувань на якість навчання та покращує обробку навчальної інформації. За результатами експертної оцінки, запропонована модель навчання показала високу ефективність, сприяє підвищенню якості викладання англійського перекладу та активності студентів [5].

Zhang J. та інші дослідники присвятили статтю удосконаленню NMT, автори відзначають, що, попри високу якість перекладу, ці моделі не відстежують явне узгодження між вхідним і вихідним реченнями. Це ускладнює інтерпретацію процесу перекладу, накладання лексичних і структурних обмежень та практичне застосування моделей у спеціалізованих доменах. Для подолання цих проблем запропоновано введення явного фразового узгодження, подібного до підходу у фразовому SMT. Розроблений алгоритм декодування дозволяє застосовувати лексичні та структурні обмеження під час генерації перекладу. Експерименти показали, що метод підвищує інтерпретованість процесу перекладу без втрати якості та значно покращує результати у завданнях із лексичними й структурними обмеженнями порівняно з традиційними NMT-системами [6].

Bhuvaneswari K. з колегами розглядали проблему обмеженості мовних ресурсів для SMT та NMT систем, які найкраще працюють для мов із великими корпусами даних. Автори пропонують гібридний підхід, що поєднує сильні сторони SMT та NMT для покращення перекладу мов із низьким ресурсом, зокрема англійської та тамільської. Початкова NMT-модель перекладає одномовний корпус і відбирає найкращі переклади для повторного навчання, при цьому SMT-таблиця фразових пар визначає оптимальні варіанти перекладу. Такий багаторазовий цикл формує великий квазіпаралельний корпус, підвищуючи ефективність моделі. Beam search-декодування дозволяє генерувати k найкращих перекладів для кожного речення. Експериментальні результати показують істотне підвищення якості: BLEU-показники *Eng→Tam* і *Tam→Eng* зросли з 11.06 і 17.06 до 19.56 і 23.49 відповідно, а оцінка METEOR підтверджує перевагу гібридної моделі над базовими NMT-системами [7].

У статті Li Z. розглянуто основні принципи роботи систем автоматичного перекладу в межах штучного інтелекту. По-перше, лінгвістичний принцип передбачає аналіз граматичної структури речення – виявлення частин мови, узгоджень і синтаксичних зв'язків. По-друге, семантичний принцип спрямований на розпізнавання значень слів у контексті, що допомагає уникати двозначностей. По-третє, статистичний або нейронний принцип базується на ймовірнісному узгодженні слів між мовами через машинне навчання. Додатково виділяється адаптивний принцип, який дає змогу системі вдосконалюватися під час використання, та гібридний

принцип, що поєднує лінгвістичні правила з нейронними моделями для підвищення точності. У сукупності ці підходи забезпечують природність і точність перекладу, роблячи його ефективним для інтелектуальних систем [8].

У дослідженні Soliman A. S. та інших дослідників були розглянуті проблеми генерації коду з природної мови. Автори описують MarianCG – трансформерну модель на базі MarianMT, яка зазвичай використовується для машинного перекладу, але у цьому випадку застосовується для перетворення описів на Python-код. Модель використовує синусоїдальне позиційне кодування для відображення позицій токенів у тексті та обходиться без нормалізації шарів. MarianCG базується на дообученні попередньо тренованої моделі перекладу, що дозволяє демонструвати, що трансформери для перекладу можна адаптувати для генерації коду. Модель тестувалася на наборах даних CoNaLa та DJANGO і показала високі результати: для CoNaLa BLEU дорівнює 34.43, точність повного збігу 10.2%, для DJANGO BLEU – 90.41, точність повного збігу 81.83%. Таким чином, MarianCG демонструє ефективність у задачі NL-to-Code, показуючи, що попередньо треновані моделі перекладу можуть бути успішно адаптовані для генерації коду з природної мови [9].

У дослідженні Demenchuk O семантичні аспекти перекладу з англійської на українську мову, що є актуальним через зростання потреби в адекватному та культурно релевантному перекладі для ефективного порозуміння у глобалізованому комунікативному середовищі. Основну увагу приділено класифікації типів семантичної відповідності, явищам семантичної дивергенції та взаємозв'язку лексичної семантики, контекстуального значення та культурно зумовлених знань, що ускладнюють процес перекладу. Дослідження показало, що семантичні відповідники часто бувають ізоморфними, проте розбіжності у лексичній структурі та ідіоматичному репертуарі потребують функційних аналогів, описових трансформацій або компенсаторних стратегій. Досягнення семантичної відповідності можливе лише за умови глибокого розуміння вихідної та цільової мов і релевантних культурних контекстів. Результати можуть бути використані для підвищення якості перекладу, розвитку міжкультурної компетентності та удосконалення моделей перекладацької діяльності [10].

У статті Han C розглядається роль машинного перекладу (MT) у сучасному глобалізованому контексті, де мовні бар'єри все ще можуть ускладнювати доступ до інформації. Автори зазначають, що іноді потребу в перекладі неможливо задовольнити виключно за допомогою людських перекладачів, тому інструменти машинного перекладу набувають популярності завдяки своєму потенціалу подолати ці труднощі. У роботі проведено систематичний огляд літератури з метою виявлення найбільш поширених систем MT, їхньої архітектури, процедур оцінки якості перекладу та визначення, які з цих систем дають найкращі результати. Дослідження фокусувалося на спеціалізованій літературі, створеній експертами з перекладу, лінгвістами та фахівцями суміжних дисциплін, що охоплює англо-іспанську мовну пару. Результати показують, що нейронний машинний переклад є домінуючим підходом у сучасному MT-сценарії, при цьому найпопулярнішою системою є Google Translate. Більшість проаналізованих робіт використовували один тип оцінки – або

автоматичну, або людську – для оцінки якості перекладу, і лише 22% робіт поєднували обидва типи оцінки. Однак понад половина робіт включала класифікацію та аналіз помилок, що є важливим аспектом для виявлення недоліків і підвищення ефективності систем машинного перекладу [11]. У статті Ponomarevskyi A. досліджується вплив контексту на варіації термінології у прикладах перекладу з англійської на українську мову на основі ключових особливостей роботи з термінологією, орієнтованою на переклад, та моделей понять у сфері штучного інтелекту [12].

Мета статті. Метою статті є теоретичне обґрунтування та структурне моделювання гібридної лінгвістично орієнтованої системи перекладу з англійської на українську мову, яка поєднує методи морфосинтаксичного аналізу з нейронними підходами автоматичного перекладу. Автори прагнули показати, що інтеграція лінгвістичних шарів у трансформерну архітектуру може суттєво підвищити якість перекладу для флективних мов, зокрема української, шляхом покращення граматичного узгодження, контекстуальної точності та природності відтворення тексту.

Виклад основного матеріалу дослідження. Дослідження теоретичних аспектів автоматизованого перекладу передбачає аналіз основних підходів, що лежать в основі створення машинних перекладачів. У сучасній лінгвістиці та комп'ютерній науці найбільш поширеними вважаються статистичні, нейронні та гібридні моделі, кожна з яких має власні переваги, обмеження та сфери застосування.

Статистичний машинний переклад (Statistical Machine Translation, SMT) ґрунтується на використанні статистичних моделей, які формуються на основі аналізу великих корпусів текстів двома мовами – мовою оригіналу (source language) та мовою перекладу (target language). Принцип роботи таких систем полягає у визначенні ймовірнісних відповідностей між одиницями вихідної мови та їхніми можливими еквівалентами у цільовій мові. Інакше кажучи, система «навчається» на основі вже перекладених текстів, виявляючи статистичні закономірності у відповідності слів, словосполучень та граматичних структур. Завдяки такому підходу статистичні системи перекладу демонструють відносно високі результати у базових завданнях, зокрема під час перекладу технічних або стандартизованих текстів, де структура речень є типовою та повторюваною. Проте основним недоліком статистичного машинного перекладу залишається обмежене врахування контексту. Оскільки система оперує лише ймовірностями появи певних мовних одиниць у паралельних корпусах, вона не здатна адекватно інтерпретувати значення слова у конкретному контексті, що призводить до семантичних помилок.

Хоча статистичний підхід і став важливим етапом розвитку машинного перекладу, його ефективність значною мірою залежить від якості та обсягу навчальних даних, а також від ступеня структурної подібності мовних пар. Ці обмеження зумовили подальший перехід дослідників до створення більш гнучких і контекстно-чутливих моделей – зокрема, нейронних систем перекладу. Сучасні дослідження показують, що врахування документного контексту, міжфразових зв'язків та термінологічної узгодженості значно підвищує якість перекладу, особливо для складних мовних пар і спеціалізованих текстів [13].

Нейронний машинний переклад (Neural Machine Translation, NMT) є сучасним підходом у сфері автоматизованого перекладу, що використовує штучні нейронні мережі для аналізу тексту та здійснення перекладу між різними мовами. На відміну від традиційних методів, NMT здатен враховувати контекст у межах речення або цілого тексту, що забезпечує більш точні та природні результати перекладу. Розвиток NMT став логічним продовженням еволюції машинного перекладу. Перші системи були побудовані на правилах та словниках, потім з'явилися статистичні моделі, які підвищили точність перекладу завдяки аналізу великих двомовних корпусів текстів. Проте статистичні системи залишалися обмеженими через недостатнє врахування семантичного та синтаксичного контексту. NMT подолав ці обмеження, використовуючи глибинні нейронні мережі, зокрема архітектури RNN (рекурентні нейронні мережі) з кодером-декодером та механізмом уваги (attention), що дозволяє системі «зосереджуватися» на релевантних частинах вхідного тексту під час генерації перекладу.

У таблиці 1 представлено порівняння основних характеристик статистичного, нейронного та гібридного перекладу.

Таблиця 1

Порівняльна характеристика систем перекладу

Підхід	Переваги	Недоліки	Типові сфери застосування
Статистичний (SMT)	Простота реалізації, швидкість	Ігнорує контекст, обмежена природність	Технічні тексти, стандартизовані документи
Нейронний (NMT)	Глибока контекстуальна обробка	Вимогливість до даних і потужностей	Багатомовний контент, локалізація
Гібридний (HMT)	Поєднує точність і гнучкість	Потребує людського втручання	Бізнес і професійна локалізація перекладів

Процес навчання NMT включає підготовку моделей на великих корпусах двомовних текстів, що дозволяє системі поступово «вчитися» відтворювати семантичні та стилістичні особливості цільової мови. Такий підхід забезпечує вищу точність перекладу, значно зменшує кількість помилок і знижує потребу в ручному редагуванні перекладів.

Тому, нейронний машинний переклад є ключовим етапом розвитку автоматизованих систем перекладу, поєднуючи штучний інтелект і лінгвістичні моделі для досягнення високої точності, природності та ефективності перекладу.

Гібридний машинний переклад (Hybrid Machine Translation, HMT) поєднує можливості автоматизованого перекладу та професійного постредагування, інтегруючи методи статистичного та нейронного перекладу з експертною оцінкою людини. Такий підхід дозволяє поєднувати швидкість і масштабованість машинного

перекладу з точністю, природністю та культурною адекватністю, забезпечуючи більш високоякісні результати, ніж використання окремих методів. Машинний компонент гібридного перекладу забезпечує оперативне та послідовне оброблення великих обсягів тексту, зменшуючи витрати та скорочуючи час на вихід матеріалу на ринок. Натомість участь професійних перекладачів гарантує відтворення стилістичних, емоційних та культурних нюансів тексту. Гібридний підхід активно застосовується провідними компаніями для локалізації контенту, міжнародної комунікації та масштабних проєктів, де одночасно важлива якість, швидкість та ефективність. Переваги гібридного перекладу включають скорочення часу на переклад, можливість швидкого виходу контенту на міжнародний ринок, високу якість та природність перекладу, адаптацію до культурних та стилістичних особливостей, підтримку багатомовного контенту та гнучкість для масштабних проєктів. Таким чином, гібридний машинний переклад є ефективним рішенням для бізнесу та професійної локалізації, поєднуючи сильні сторони технологій автоматизованого перекладу та людського досвіду.

Машинний переклад пройшов кілька етапів розвитку, що визначають різні підходи до обробки тексту. Статистичний машинний переклад (SMT) базується на аналізі двомовних корпусів текстів і визначенні ймовірнісних відповідностей між словами та фразами. Він ефективний для базових текстів, але обмежено враховує контекст, що може призводити до помилок. Нейронний машинний переклад (NMT) використовує глибинні нейронні мережі, кодери-декодери та механізми уваги, що дозволяє враховувати контекст на рівні речення або тексту, підвищуючи точність і природність перекладу. Гібридний машинний переклад (HMT) поєднує машинні методи та професійне постредагування, забезпечуючи швидкість і масштабованість машинного перекладу разом із точністю, стилістичною адекватністю та врахуванням культурних особливостей, що робить його оптимальним для складних і масштабних проєктів.

Англійська та українська мови відрізняються на рівні морфології та синтаксису, що створює складнощі для систем автоматизованого перекладу грамотно передати сенс тексту. Українська мова характеризується розвиненою морфологічною системою: наявністю відмінків, граматичного роду, числа, аспекту дієслова та узгодження слів у реченні. Це дозволяє гнучко змінювати порядок слів для передачі семантичного наголосу, що підкреслює інформаційно важливі елементи речення. Англійська мова, навпаки, має спрощену морфологію та більш фіксований синтаксичний порядок, де зміна порядку слів може змінювати значення речення або робити його граматично неправильним.

Особливої уваги потребують конструкції, які виконують нетипові граматичні функції. Герундій (Gerund) та інфінітив (Infinitive) у англійській мові можуть виступати одночасно як: підмет, додаток або означення, тоді як в українській мові для передачі тих самих значень часто використовуються дієслівні конструкції з інфінітивом або іменники, що потребує перестановки та адаптації синтаксису. Складні речення з наголосом (Cleft sentences) демонструють ще одну асиметрію: англійські структури для логічного виділення інформації не завжди мають прямий

український еквівалент, і під час перекладу необхідно зберегти семантичний акцент, переставляючи члени речення.

Такі лінгвістичні відмінності створюють додаткові виклики для систем автоматизованого перекладу. Машини повинні не лише зіставляти слова і словосполучення, а й правильно передавати граматичну узгодженість, семантичну точність і стилістичні нюанси тексту. Збереження цих аспектів критично важливе для природності перекладу та адекватного сприйняття інформації носіями цільової мови. Врахування морфологічних та синтаксичних особливостей англійської та української мов є необхідною умовою розробки ефективних нейронних та гібридних систем машинного перекладу, що здатні обробляти складні текстові структури та підтримувати високий рівень якості перекладу.

При перекладі з англійської на українську системи машинного перекладу часто стикаються з типовими помилками. Однією з основних проблем є некоректне відтворення порядку слів. Англійська мова має фіксований порядок підмет-присудок-додаток, тоді як українська допускає більшу свободу синтаксичного розташування для підкреслення смислового наголосу. Машини часто застосовують прямий порядок слів, що призводить до втрати основної думки.

Ще однією частою помилкою є неврахування контексту при перекладі конструкцій Gerund і Infinitive. Герундій англійської мови часто передається як іменник або дієслівна конструкція в українській, і прямий переклад слова у слово може створювати граматично неправильне речення. Інфінітиви англійської мови також потребують адаптації під український синтаксис, і без правильної перестановки слів та узгодження з підметом переклад може втратити точність значення.

Складність становлять і Cleft sentences, які використовуються для логічного виділення інформації. Прямий переклад таких речень часто спотворює наголос і семантичну структуру, оскільки українська мова передає логічний акцент іншими засобами, наприклад перестановкою слів або використанням частки «саме».

На рисунку 1 подано схему, що ілюструє головні морфологічні відмінності між англійською та українською мовами, які створюють труднощі для систем автоматичного перекладу.

Іншими типовими помилками є некоректне узгодження родів і чисел, невірний переклад часових форм дієслів та неадекватне відтворення фразових дієслів, що не мають прямого українського еквівалента. Всі ці труднощі підкреслюють необхідність застосування систем перекладу, здатних враховувати граматичну, семантичну та стилістичну структуру тексту для досягнення природності та точності перекладу.

Семантичне узгодження в українській мові є важливою складовою побудови граматично правильного, об’єктивного та точного речення. Воно передбачає взаємозв’язок між елементами речення на рівні значення, коли граматичні форми підлягають логічному та смисловому узгодженню. Зокрема, підмет і присудок, визначення та іменник, числівник та іменник повинні відповідати один одному не лише за граматичними категоріями, але й за семантичним змістом. Наприклад, в українській мові категорія роду та числа тісно пов’язана з семантикою іменника:

для живих істот граматичний рід часто збігається з біологічним, тоді як для неживих предметів рід визначається морфологічними правилами. Крім того, словосполучення, що включають прикметники, дієприкметники або числівники, підпорядковуються принципу семантичної сумісності: значення слів повинно гармонійно поєднуватися, щоб речення передавало правильну інформацію.

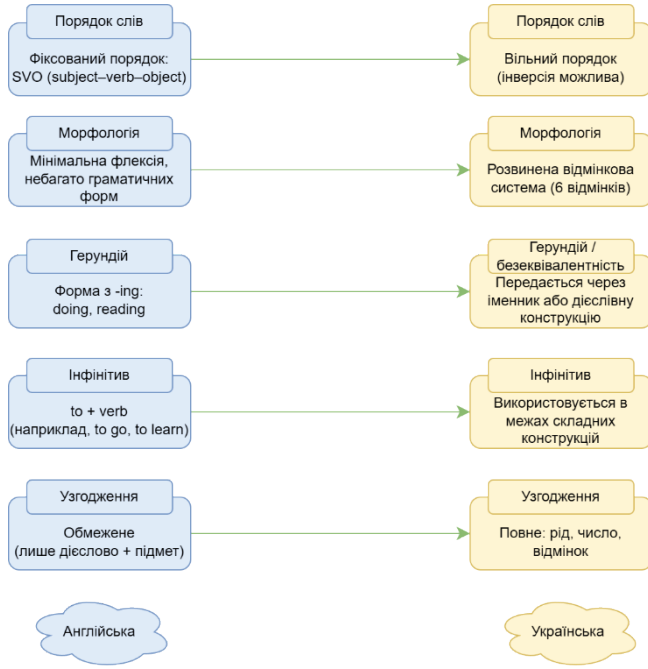

Рис. 1. Морфологічні відмінності англо-українського перекладу

Особливо важливо звертати увагу на семантичне узгодження в контексті перекладу з мов із менш розвиненою морфологією. Дослівне відтворення структури речення часто призводить до порушення узгодження та втрати логічності тексту. Тому при автоматизованому перекладі системи повинні враховувати не лише граматичну коректність, а й семантичну взаємозв'язність елементів речення, щоб досягти природності та точності українського тексту.

Мовознавці відзначають, що англо-український переклад пов'язаний із низкою складних лінгвістичних викликів, обумовлених відмінностями граматичних категорій числа, роду та виду. За словами дослідників, успішний переклад можливий лише за умови врахування граматичних відмінностей для забезпечення точності, природності та стилістичної адекватності українського тексту.

Англійські ідіоми часто становлять суттєву проблему для перекладу на українську мову через їхню образність та культурну специфіку. Фразеологізми на кшталт "piece of cake" або "break the ice" мають переносне значення, яке не можна передати дослівно українським перекладом. Відзначається, що для адекватного відтворення значення таких конструкцій необхідно застосовувати адаптований переклад,

який передає смисл та функцію ідіоми в контексті, наприклад “дуже легко” або “розтопити лід у спілкуванні”, а не дослівний переклад кожного слова. Цей аспект підкреслює важливість контексту та культурної обізнаності для точності і природності перекладу.

Навпаки, переклад українських фразеологізмів на англійську мову також є складним завданням через відмінності у культурному та мовному середовищі. Фразеологізми, такі як “водити за ніс”, “ставити крапку над і” або “як сіль на рану”, не мають буквальних англійських аналогів, і прямий переклад може бути незрозумілим або створювати хибне значення. Для забезпечення адекватного сприйняття в англомовному середовищі переклад часто здійснюють через адаптовані ідіоматичні відповідники, наприклад “to lead by the nose”, “to dot the i’s and cross the t’s”, “to add insult to injury”, що дозволяє зберегти образність та функціональне навантаження українських висловів.

Англійська лексика відзначається високим ступенем полісемії, що створює суттєві труднощі для вибору адекватних українських еквівалентів. Одне і те саме слово може мати кілька значень залежно від контексту: наприклад, “bank” може означати як «банк» у фінансовому сенсі, так і «берег» річки. Мовознавці підкреслюють, що машинний переклад без врахування контексту часто призводить до семантичних помилок, оскільки системи не можуть самостійно визначити правильний еквівалент. Для точного перекладу необхідно використовувати алгоритми, здатні аналізувати контекст на рівні речення або абзацу та враховувати тематичну спрямованість тексту, що дозволяє обрати найбільш відповідне значення лексеми.

Подібні труднощі виникають через контекстну залежність синтаксичних конструкцій англійської мови. Багато структур, таких як умовні речення, пасивні конструкції або інверсії, можуть змінювати значення висловлювання залежно від контексту, в якому вони використовуються. Українська мова має більшу свободу порядку слів і більш розвинену систему морфологічного узгодження, що вимагає від перекладача або системи машинного перекладу адаптувати синтаксичну структуру, щоб зберегти семантичний наголос і точність повідомлення.

У сучасній практиці автоматизованого перекладу важливе місце посідають програмні рішення, що реалізують різні архітектурні підходи до обробки природної мови. Такі системи, як: Google Translate, DeepL, Meta AI Translator та Grammarly, є прикладами високотехнологічних платформ, які поєднують алгоритми глибокого навчання, нейронні мережі та мовні моделі нового покоління. Вони не лише автоматизують процес перекладу, а й забезпечують адаптацію до контексту, стилю та мовних норм. Кожна з цих платформ має власну архітектуру: Google Translate базується на системі Neural Machine Translation (GNMT), DeepL використовує вдосконалені багаторівневі трансформерні моделі, Meta AI Translator має впроваджені масштабовані багатомовні мережі, а Grammarly застосовує гібридний підхід, поєднуючи граматичний аналіз із контекстною семантикою.

Робота сучасних систем автоматизованого перекладу з українською мовою має низку специфічних особливостей, зумовлених як структурною складністю самої мови, так і обмеженою кількістю високоякісних паралельних корпусів.

Українська мова належить до флективних мов із розвиненою морфологічною системою, що включає шість відмінків, категорію роду та складну систему дієвідмінювання. Для більшості нейронних моделей, зокрема Google Translate чи DeepL, така варіативність створює труднощі під час точного узгодження форм у контексті речення. Крім того, на відміну від англійської, українська характеризується вільним порядком слів і наявністю лексичних та стилістичних відтінків, які не завжди мають прямі еквіваленти. Через це системи перекладу часто помиляються в передачі аспектуальної семантики, граматичного роду чи комунікативного наголосу. Останні дослідження демонструють, що якість перекладу українських текстів суттєво зростає при використанні трансформерних моделей, навчених на локалізованих корпусах, зокрема в рамках проектів Meta NLLB (No Language Left Behind) та OPUS-MT, які забезпечують глибшу контекстуалізацію і точніше відтворення морфосинтаксичних зв'язків.

Попри значні досягнення в галузі нейронного машинного перекладу, застосування моделей BERT, MarianMT та подібних архітектур у контексті англо-українського перекладу має певні обмеження. Модель BERT (Bidirectional Encoder Representations from Transformers), створена для завдань розуміння мови, а не генерації тексту, демонструє високу ефективність у контекстному аналізі, проте не призначена для повноцінного перекладу речень. Її використання можливе лише у гібридних системах, де BERT виконує функцію семантичного аналізатора перед передачею даних у генеративну модель. Натомість MarianMT, спеціально розроблена для багатомовного перекладу в межах проєкту OPUS, забезпечує якісну двосторонню трансформацію тексту між англійською та українською, проте її ефективність знижується при обробці складних синтаксичних структур або стилістично насичених текстів.

Процес машинного перекладу сучасного зразка ґрунтується на багаторівневій обробці природної мови, у якій поєднуються лінгвістичні, статистичні й нейронні методи. Проте традиційні моделі перекладу, орієнтовані лише на текстову послідовність, часто залишають поза увагою глибинні граматичні взаємозв'язки, що є критично важливими для морфологічно складних мов, зокрема української. З огляду на це запропоновано гібридну лінгвістично орієнтовану модель перекладу, що інтегрує попередній морфосинтаксичний аналіз у типову трансформерну архітектуру.

Основна концепція моделі полягає у тому, що англомовний текст спочатку проходить етап морфосинтаксичної нормалізації – токенізацію, визначення частин мови, лематизацію й аналіз граматичних залежностей. Цей блок виконує функцію попереднього лінгвістичного фільтра, який виділяє ключові структурні та граматичні параметри речення. Далі формується вектор ознак, що відображає морфологічні характеристики – рід, число, відмінок, вид, час і синтаксичні зв'язки між словами. Отримані лінгвістичні ознаки інтегруються у нейронний модуль перекладу, заснований на архітектурі типу Transformer.

На етапі кодування (encoder) конструюється багатовимірне представлення (embedding), яке поєднує лексичний та граматичний контекст. У механізмі уваги

(attention layer) здійснюється селекція найрелевантніших зв'язків між словами, що дозволяє врахувати контекст не лише на рівні речення, а й дискурсу. У декодері відбувається генерація послідовності українських токенів, які передають зміст вихідного тексту із збереженням морфологічної, синтаксичної та семантичної узгодженості.

Далі застосовується модуль морфогенерації, який адаптує згенерований текст під граматичні правила української мови – узгоджує відмінки, рід, число, вид і порядок слів. Завершальний етап – постредагування, що включає автоматичну граматичну та стилістичну корекцію, а також оцінювання якості перекладу.

Для кількісної оцінки ефективності моделі використовуються міжнародні метрики BLEU (Bilingual Evaluation Understudy) та TER (Translation Edit Rate). Метрика BLEU відображає схожість зразка машинного перекладу з еталонним перекладом людини на основі n-грам статистики. Вищі значення BLEU свідчать про лексико-синтаксичну близькість моделі до людського перекладу. Натомість метрика TER визначає кількість редагувань (вставлення, видалення, заміна, перестановка), необхідних для перетворення машинного перекладу на референтний текст, тобто вимірює «вартість редагування». Таким чином, зменшення TER і збільшення BLEU є показниками підвищення якості моделі.

Метрика BLEU (Bilingual Evaluation Understudy) розраховується на основі збігу n-грам між машинним перекладом (candidate) і референтним (людським) перекладом (reference). Формула має вигляд:

$$BLEU = BP \cdot \exp\left(\sum_{i=1}^{N} w_i \cdot \ln\left(p_i \right) \right), \tag{1}$$

де *BP – brevity penalty* (штраф за надмірно короткий переклад), що визначається як:

$$BP = \begin{cases} 1, & якщо\ c > r \\ e^{(1-r/c)}, & якщо\ c \le r \end{cases}$$

- довжина референтного перекладу,
- довжина машинного перекладу,
- максимальний розмір *n*-грам (зазвичай 4),
- вагові коефіцієнти для *i*-грам (зазвичай рівні, тобто $w_i = \frac{1}{N}$),
- модифікована точність *i*-грам, обчислена як

$$p_i = \frac{\sum_{ngram \in Cand} \min\left(Count_{cand}\left(ngram\right), Count_{ref}\left(ngram\right) \right)}{\sum_{ngram \in Cand} Count_{cand}\left(ngram\right)}. \tag{2}$$

Отже, BLEU-оцінка є експоненціально зваженим середнім модифікованих точностей n-грам і враховує штраф за коротші тексти. Показник варіюється від 0 до 1: значення, близькі до 1, означають високу схожість із людським перекладом.

Метрика TER (Translation Edit Rate) визначає, наскільки часто машинний переклад потребує редагування, щоб стати ідентичним до референтного. Формула TER має такий вигляд:

$$TER = \frac{E}{R}, \tag{3}$$

де E – кількість елементарних редагувань, необхідних для перетворення машинного перекладу у референтний (включаючи вставлення, видалення, заміну та перестановку слів); R – загальна кількість слів у референтному перекладі.

Низьке значення TER свідчить про меншу кількість необхідних редагувань і, відповідно, про вищу якість перекладу. У наукових дослідженнях вважається, що системи з BLEU > 0.5 та TER < 0.3 демонструють продуктивність, близьку до рівня людського перекладача.

Такі формули дозволяють кількісно оцінити ефективність алгоритмів машинного перекладу та порівняти різні архітектури систем, включно з гібридними моделями, що інтегрують морфологічний аналіз і нейронні методи.

Загалом робота запропонованої архітектури полягає у послідовній інтеграції лінгвістично обґрунтованого аналізу та нейронного навчання. Це дозволяє моделі не лише перекладати текст, а й відтворювати внутрішні граматичні зв'язки оригіналу, що є надзвичайно важливим для української мови з її високим рівнем флексійності.

На рисунку 2 подано узагальнену структурну схему, що демонструє повний цикл обробки – від англомовного вхідного тексту до морфологічно й стилістично коректного українського перекладу, отриманого в результаті взаємодії лінгвістичних та нейронних компонентів системи.

Основними чинниками, що істотно впливають на якість англо-українського машинного перекладу, залишаються висока морфологічна варіативність української мови, обмежена наявність якісних паралельних корпусів для навчання моделей і складна багаторівнева контекстна неоднозначність англійських синтаксичних конструкцій. Брак великих та збалансованих корпусів текстів призводить до нестійкості результатів перекладу, особливо тоді, коли система стикається з рідковживаними формами, ідіомами чи нестандартними граматичними структурами. У свою чергу, морфологічна багатошаровість української мови – наявність відмінкових форм, категорій роду, числа, виду, виду дієслів та узгодження – створює для алгоритмів додаткові труднощі, які не властиві аналітичним мовам, наприклад англійській. Ці фактори зумовлюють потребу у збільшенні обсягу навчальних даних, систематизації корпусних джерел і більш глибокому урахуванні лінгвістичних закономірностей при моделюванні перекладацьких процесів.

Такі виклики підкреслюють необхідність переходу від універсальних нейронних моделей до лінгвістично орієнтованих гібридних рішень, що поєднують алгоритмічний аналіз контексту з попереднім морфосинтаксичним опрацюванням тексту. Інтеграція морфологічних, синтаксичних та семантичних ознак на перед-нейронному етапі забезпечує ширший спектр аналітичних можливостей системи, дозволяє коректно узгоджувати граматичні форми, уникати семантичних спотворень і значно підвищує природність українського перекладу. Такий підхід дає змогу створити гнучку адаптивну архітектуру, здатну працювати з різними жанрами текстів – від науково-технічних до образних чи публіцистичних – і краще передавати мовну динаміку між англійським аналітизмом і українською флексійною структурою.

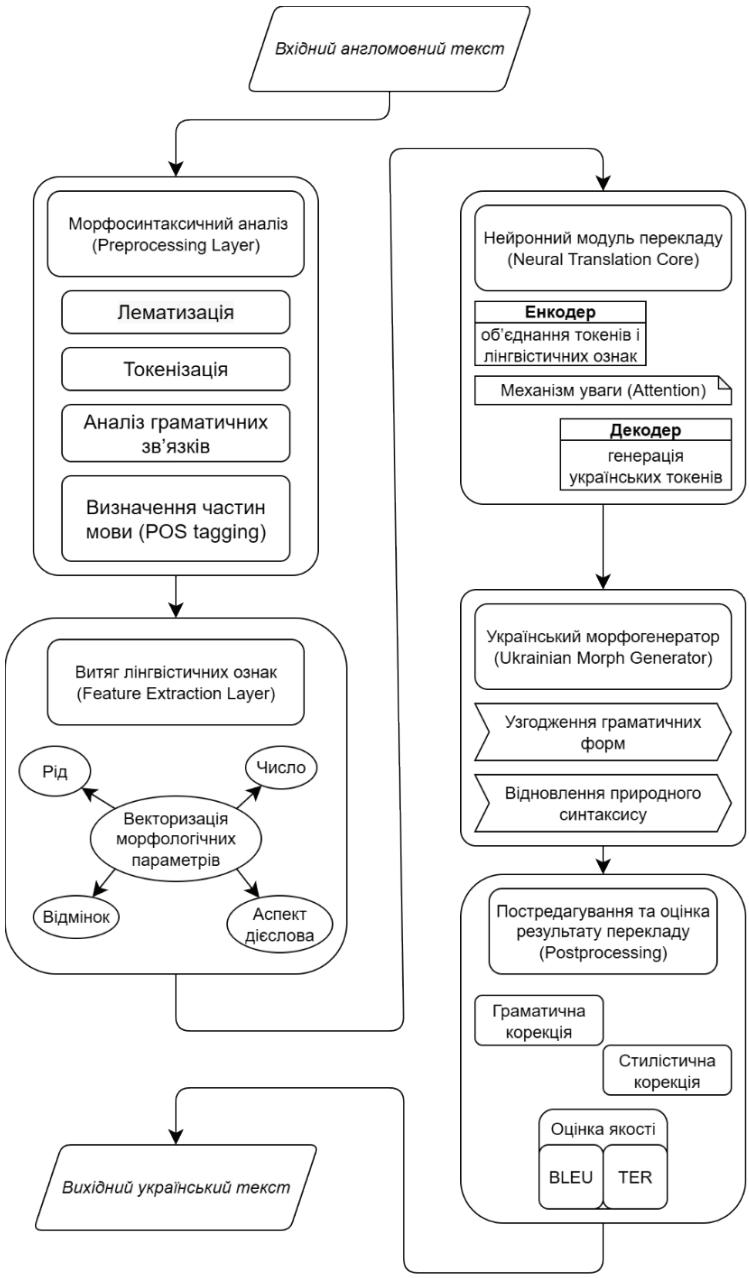

Рис. 2. Гібридна лінгвістично-орієнтована модель перекладу

Висновки. У результаті проведеного аналізу виявлено, що традиційні нейронні моделі перекладу демонструють ефективність у синтаксично простих мовах, проте втрачають точність при роботі з українською мовою через морфологічну складність, вільний порядок слів та систему узгодження роду, числа й відмінків.

Запропонована гібридна лінгвістично-орієнтована модель перекладу дозволяє компенсувати ці недоліки шляхом включення попереднього етапу морфосинтаксичного аналізу перед нейронною обробкою тексту.

Структурна архітектура моделі складається з п'яти взаємопов'язаних компонентів: блоку морфосинтаксичного аналізу, модуля витягу ознак, нейронного трансформерного ядра, морфогенератора та підсистеми постредагування. Застосування метрик BLEU і TER у межах фінального етапу оцінювання створює можливість об'єктивно порівнювати якість перекладу з еталонними людськими результатами.

Подальші дослідження можуть бути спрямовані на розробку експериментального корпусу англо-українських текстів для тренування та тестування запропонованої моделі, розширення її функціональних можливостей до багатомовного перекладу, а також на інтеграцію підходу до прикладних середовищ – офісних додатків, освітніх систем і гуманітарних перекладацьких сервісів. Очікується, що практична реалізація цієї архітектури сприятиме підвищенню якості україномовного контенту у цифровому просторі та розширить перспективи автоматизованої обробки природної мови.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tomenchuk M., Popovych K. Large Language Models and Machine Translation. *Věda a Perspektivy*. 2024. T. 11. C. 422–431. DOI: 10.52058/2695-1592-2024-11(42)-422-431.
2. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. *Mathematics*. 2023. T. 11, № 4. C. 904. DOI: 10.3390/math11040904.
3. Mandziy K. S., Yurlova U. V., Dilai M. P. English-Ukrainian Parallel Corpus of IT Texts: Application in Translation Studies. In: *COLINS*. 2022. C. 724–736. URL: https://api.semanticscholar.org/CorpusID:250625296.
4. Vysotska V. Computer linguistic system modelling for Ukrainian language processing. In: *CEUR Workshop Proceedings*. 2024. T. 3722. C. 288–342. URL: https://dblp.org/rec/conf/colins/Vysotska24b.bib.
5. Che C. Application of Differential Evolution Algorithm in the Construction and Simulation of Interactive English Translation Teaching Mode. *Advances in Multimedia*. 2022. T. 2022, № 1. C. 6182551. DOI: 10.1155/2022/6182551.
6. Zhang J., Luan H., Sun M., Zhai F., Xu J., Liu Y. Neural machine translation with explicit phrase alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021. T. 29. C. 1001–1010. DOI: 10.1109/TASLP.2021.3057831.
7. Bhuvaneswari K., Varalakshmi M. Efficient incremental training using a novel NMT-SMT hybrid framework for translation of low-resource languages. *Frontiers in Artificial Intelligence*. 2024. T. 7. C. 1381290. DOI: 10.3389/frai.2024.1381290.
8. Li Z., Gao S., Li X., Asghari H. B. Strengthening Research of SYSTRAN in the Field of Artificial Intelligence Automatic Translation. In: International Conference on Applications and Techniques in Cyber Intelligence. *Cham: Springer International Publishing*. 2022. C. 626–633. DOI: 10.1007/978-3-031-29097-8_74.

9. Soliman A. S., Hadhoud M. M., Shaheen S. I. MarianCG: a code generation transformer model inspired by machine translation. *Journal of Engineering and Applied Science*. 2022. Т. 69, № 1. С. 104. DOI: 10.1186/s44147-022-00159-4.
10. Demenchuk O. Semantic aspects of translation: English-Ukrainian correspondences. *Наукові записки Вінницького державного педагогічного університету імені Михайла Коцюбинського. Серія: Філологія (мовознавство)*. 2025. № 40. С. 141–147. DOI: 10.31652/2521-1307-2025-40-15.
11. Han C. Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): *Proposing a unifying framework for research. Interpreting and Society*. 2025. Т. 5, № 1. С. 27–55. DOI: 10.1177/27523810251322645.
12. Ponomarevskyi A. Variability of artificial intelligence related terms used in translation from English into Ukrainian [Doctoral dissertation, Kauno technologijos universitetas]. *Laisvai prieinamas internete / Unrestricted online access*. 2022. URL: https://epubl.ktu.edu/object/elaba:132032271/.
13. Castilho S., Knowles R. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*. 2025. Т. 31, № 4. С. 986–1016. DOI: 10.1017/nlp.2024.7.

REFERENCES

1. Tomenchuk, M., & Popovych, K. (2024). Large Language Models and Machine Translation. Věda a Perspektivy. 11. 422–431. https://doi.org/10.52058/2695-1592-2024-11(42)-422-431.
2. Lytvyn, V., Pukach, P., Vysotska, V., Vovk, M., & Kholodna, N. (2023). Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. Mathematics, 11(4), 904. https://doi.org/10.3390/math11040904.
3. Mandziy, K. S., Yurlova, U. V., & Dilai, M. P. (2022). English-Ukrainian Parallel Corpus of IT Texts: Application in Translation Studies. In COLINS. 724-736. https://api.semanticscholar.org/CorpusID:250625296.
4. Vysotska, V. (2024). Computer linguistic system modelling for Ukrainian language processing. In CEUR Workshop Proceedings. 3722. 288-342. https://dblp.org/rec/conf/colins/Vysotska24b.bib.
5. Che, C. (2022). Application of Differential Evolution Algorithm in the Construction and Simulation of Interactive English Translation Teaching Mode. *Advances in Multimedia*, 2022(1), 6182551. https://doi.org/10.1155/2022/6182551.
6. Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., & Liu, Y. (2021). Neural machine translation with explicit phrase alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1001-1010. https://doi.org/10.1109/TASLP.2021.3057831.
7. Bhuvaneswari, K., & Varalakshmi, M. (2024). Efficient incremental training using a novel NMT-SMT hybrid framework for translation of low-resource languages. *Frontiers in Artificial Intelligence*, 7, 1381290. https://doi.org/10.3389/frai.2024.1381290.
8. Li, Z., Gao, S., Li, X., & Asghari, H. B. (2022, June). Strengthening Research of SYSTRAN in the Field of Artificial Intelligence Automatic Translation. In *International Conference on Applications and Techniques in Cyber Intelligence* (pp. 626-633). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-29097-8_74.

9. Soliman, A. S., Hadhoud, M. M., & Shaheen, S. I. (2022). MarianCG: a code generation transformer model inspired by machine translation. *Journal of Engineering and Applied Science*, 69(1), 104. https://doi.org/10.1186/s44147-022-00159-4.
10. Demenchuk, O. (2025). Semantic aspects of translation: English-Ukrainian correspondences. *Naukovi zapysky Vinnyts'koho derzhavnoho pedahohichnoho universytetu imeni Mykhayla Kotsyubyns'koho. Seriya: Filolohiya (movoznavstvo), (40)*, (40), 141-147. https://doi.org/10.31652/2521-1307-2025-40-15.
11. Han, C. (2025). Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): Proposing a unifying framework for research. *Interpreting and Society*, 5(1), 27-55. https://doi.org/10.1177/27523810251322645.
12. Ponomarevskyi, A. (2022). *Variability of artificial intelligence related terms used in translation from English into Ukrainian* [Doctoral dissertation, Kauno technologijos universitetas]. Laisvai prieinamas internete / Unrestricted online access. https://epubl.ktu.edu/object/elaba:132032271/.
13. Castilho, S., & Knowles, R. (2025). A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4), 986-1016. https://doi.org/10.1017/nlp.2024.7.

doi: 10.32403/0554-4866-2025-2-90-128-145

HYBRID LINGUISTICALLY ORIENTED MODEL OF ENGLISH–UKRAINIAN MACHINE TRANSLATION

I. M. Liakh, M. P. Klyap, A. Yu. Tsipino, A. M. Roman

Uzhhorod National University, Faculty of Information Technology,
3 Narodna Square, Uzhhorod, 88000, Ukraine

The article explores the evolutionary development of machine translation – from rule-based systems (RBMT) and statistical methods (SMT) to next-generation neural architectures (NMT). Particular attention is devoted to the comparative analysis of morphological, syntactic, and semantic distinctions between English and Ukrainian that determine the limitations of automated translation accuracy. Among the principal divergences identified are inflectional variability, differences in word order, grammatical gender, number, aspect, and agreement categories, all of which constitute linguistic barriers to the machine reproduction of natural Ukrainian syntax.

To address these challenges, a hybrid linguistically oriented translation model is proposed, integrating morphosyntactic analysis mechanisms with a neural network architecture. The article provides a detailed description of the system's stepwise framework, which comprises a morphosyntactic normalization preprocessing block, a linguistic feature extraction module, a neural translation core, a Ukrainian morphogenerator, and a post-editing subsystem. For an objective evaluation of the proposed architecture's performance, the implementation of international translation quality metrics – BLEU

(Bilingual Evaluation Understudy) and TER (Translation Edit Rate) – is envisaged, enabling the assessment of translation accuracy and naturalness relative to human-generated reference outputs.

The authors substantiate the feasibility of incorporating grammatical and contextual rules as integral components of neural models designed for the translation of inflectional languages, particularly Ukrainian. This approach is aimed at enhancing the accuracy, stylistic naturalness, and contextual adequacy of computational translation systems. The research findings outline prospects for further development in constructing large-scale English–Ukrainian parallel corpora, experimentally validating the proposed model, and extending the approach to multilingual translation platforms. In practical terms, the study contributes to the advancement of high-quality Ukrainian-language translation tools integrated into contemporary natural language processing systems.

Keywords: *automated translation; machine translation; neural network; morpho-syntactic analysis; BLEU; TER.*

Стаття надійшла до редакції 04.09.2025.
Received 04.09.2025.