Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мережних технологій факультету інформатики

ОГЛЯД ОСНОВНИХ ПРОБЛЕМ NLP ДЛЯ УКРАЇНСЬКОЇ МОВИ, ЩО СТРИМУЮТЬ РОЗВИТОК ГАЛУЗІ / PROBLEMS IN UKRAINIAN NLP

Курсова робота

за спеціальністю “Комп’ютерні науки” 122

Керівник курсової роботи

д.т.н., проф. Глибовець А.М.

_____

(підпис)

“____” _____ 2025 р.

Виконала студентка 4-го року навчання,

Освітньої програми “Комп’ютерні науки”, 122

Мудра Катерина Володимирівна

“____” _____ 2025 р.

Київ 2025

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мережних технологій факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри мережних технологій

проф., д.ф.-м.н.

_____ Г.І. Малашонок

(підпис)

“____”_____ 2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студентці Мудрій Катерині Володимирівні 4-го року навчання факультету інформатики

ТЕМА: ОГЛЯД ОСНОВНИХ ПРОБЛЕМ NLP ДЛЯ УКРАЇНСЬКОЇ МОВИ, ЩО СТРИМУЮТЬ РОЗВИТОК ГАЛУЗІ

Зміст курсової роботи:

Індивідуальне завдання
Календарний план
Анотація
Вступ
Розділ 1. Огляд NLP
Розділ 2. Стан NLP в Україні
Розділ 3. Проблеми в українському NLP
Висновки
Список використаної літератури

Дата видачі “_____”_____ 2025 р. Керівник

(підпис)

Завдання отримав

(підпис)

Календарний план виконання курсової роботи

Тема: Огляд основних проблем NLP для української мови, що стримують розвиток галузі

№ з/п	Назва етапу	Термін виконання	Примітка
1	Постановка задачі курсової роботи	Лютий 2025	
2	Дослідження історії та теорії NLP	Березень 2025	
3	Огляд технологій	Квітень 2025	
4	Виконання практичної частини	Квітень 2025	
5	Оформлення текстової частини	Квітень 2025	
6	Висновок	Квітень 2025	

Студентка Мудра К.В.

Керівник Глибовець А.М.

“___”_____2025 р.

ЗМІСТ

АНОТАЦІЯ

У курсовій роботі досліджено основи обробки природної мови (NLP) як важливої складової штучного інтелекту, зокрема в контексті української мови. Охарактеризовано основні напрямки NLP, зокрема Natural Language Understanding та Natural Language Generation, а також їх застосування у різних сферах. Розглянуто історію розвитку NLP в Україні та сучасні досягнення у створенні лінгвістичних ресурсів і моделей для української мови. Окремо акцентовано на специфічних лінгвістичних та технічних викликах, з якими стикається обробка української мови, таких як багатозначність та контекстна неоднозначність. У роботі також запропоновано шляхи вирішення цих проблем, що відкриває нові можливості для розвитку технологій штучного інтелекту в Україні.

Ключові слова: обробка природної мови, NLP, штучний інтелект, українська мова, лінгвістичні ресурси, аналіз, трансформери, багатозначність, машинний переклад, контекстна неоднозначність, корпус.

ВСТУП

Обробка природної мови (NLP) є однією з найбільш перспективних та інноваційних галузей штучного інтелекту, яка сприяє розвитку технологій, здатних здійснювати взаємодію між комп'ютерами та людською мовою. Задачі NLP охоплюють широкий спектр від базових операцій з текстами до складних систем розпізнавання мови та генерації контенту. Вони включають машинний переклад, автоматичний аналіз емоцій, створення діалогових систем та багато інших застосувань, що змінюють способи взаємодії з інформацією у сучасному світі.

Мета завдання: Основною метою цієї курсової роботи є дослідження основ обробки природної мови (NLP) з фокусом на застосування цієї технології для обробки української мови. Зокрема, робота має на меті аналіз наявних лінгвістичних ресурсів, проблем, що виникають при обробці українського тексту, а також шляхів вирішення цих проблем для подальшого вдосконалення NLP-систем.

Актуальність роботи: На сьогодні існують значні виклики у створенні точних і ефективних моделей для обробки українського тексту, що зумовлено її лінгвістичними особливостями та недостатньою кількістю якісних ресурсів. Тому, дослідження в цій галузі має не тільки теоретичне, але й практичне значення для розвитку технологій штучного інтелекту в Україні, зокрема у сферах автоматичного перекладу, аналізу тексту, створення інтелектуальних систем та ін.

Перший розділ присвячений основам NLP, зокрема концепціям, що охоплюють Natural Language Understanding (NLU) і Natural Language Generation (NLG), а також основним напрямкам і застосуванню цих технологій у різних сферах, таких як медицина, фінанси та право.

Другий розділ висвітлює історію розвитку NLP в Україні, починаючи з наукових досягнень Тараса Вінцюка в 1960-х роках і до сучасних успіхів у створенні лінгвістичних ресурсів та моделей для української мови.

У третьому розділі аналізуються специфічні проблеми, з якими стикається обробка української мови, зокрема багатозначність, контекстна неоднозначність та труднощі у створенні репрезентативних корпусів для навчання мовних моделей. Крім того, обговорюються шляхи подолання цих проблем та можливості для подальшого розвитку NLP в Україні.

РОЗДІЛ 1. ОГЛЯД NLP

1.1. Визначення NLP

NLP (Natural Language Processing, укр. – обробка природньої мови) – це підгалузь штучного інтелекту, яка являє собою комп'ютеризований підхід до розпізнання, обробки та аналізу людської мови. [1] Людська мова, у свою чергу, - це мова, яка розвинулась природнім чином як спосіб комунікації між людьми і яка не була створена спеціально, як, наприклад, мови програмування для комп'ютерів. [2] Подібно до того, як люди в природі комунікують одне з одним, постала потреба комунікації людини з комп'ютером. Враховуючи той момент, що нам, як людям, складно вивчити на пам'ять машинно-орієнтовану мову з такою легкістю, як ми вчимо, наприклад, англійську чи німецьку, дослідники у галузі інформаційних наук почали дедалі глибше розвивати сферу NLP, аби не ми «розуміли» комп'ютер, а комп'ютер «розумів» нас. [3]

NLP як природню обробку мов дуже часто плутають із іншими поняттями. Варто розрізняти NLP (Natural Language Processing), NLP (Neurolinguistic Programming), CL (Computational Linguistic) та NLU (Natural Language Understanding):

Natural Language Processing – царина штучного інтелекту, яка допомагає комп'ютерам розуміти природню мову (зокрема сприймати та обробляти текстові, аудіо та відео, де присутня природня мова) та відтворювати її (наприклад, генерація тексту, аудіо, відео)

Neurolinguistic Programming – напрям у психології (поки що не є 100% доказовим), технологія, яка за допомогою різних прийомів може змінювати поведінку та світогляд людини через «програмування» її мислення та думок

Computational Linguistic – підгалузь лінгвістики, яка досліджує комп'ютерні моделі природніх мов. [4] Комп'ютерна (обчислювальна лінгвістика) більше орієнтована на те, як комп'ютер може автоматично аналізувати мовні конструкції: надавати семантичний чи граматичний розбір речення, визначати лінгвістичні закономірності тощо.

NLP також можна поділити на дві такі складові:

- NLU (Natural Language Understanding) – розуміння природньої мови, підвид природньої обробки мови. Використовує синтаксичний та семантичний аналіз тексту для побудови онтології речень (зв'язки між окремими словами та фразами), щоб «зрозуміти» суть речення.

- NLG (Natural Language Generation) – відтворення природньої мови, підвид природньої обробки мови. На відміну від NLU, сфокусоване на генерації вихідного тексту (письмового чи усного) на основі вхідних даних. Для досягнення цього використовується морфологічний, лексичний, синтаксичний та семантичний аналіз. [5]

NLP галузь не могла б розвиватись без великих зусиль працівників цієї сфери. NLP інженери є сьогодні одними з найпрестижніших та найзатребуваніших працівників у галузі інформаційних технологій. Спеціалісти NLP працюють у різних сферах, наприклад, фінанси, охорона здоров'я, страхування, юриспруденція тощо. [6]

1.2. Основні задачі NLP

Фактично NLP заснований на двох частинах: множині різних теорій та множині технологій. [1] Кожна з цих множин покликана розв'язати ту чи іншу проблему у галузі NLP. Зокрема різні розробки NLP надають такі переваги:

- Автоматизація рутинних задач. Там, де людина виконувала одноманітну нецікаву та ручну роботу, зараз застосовуються різноманітні інструменти NLP. Зокрема йдеться про використання чатботів, генерація документів на основі вхідних даних та попередньо розроблених шаблонів, видобування даних із тексту, класифікація документів тощо. Особливо це є потрібним працівникам сфери обслуговування клієнтів, юристам тощо. [6]

- Покращений аналіз даних. На прикладі усім відомого ChatGPT стало зрозуміло, що мовні моделі можуть виконувати роль асистента в задачах, які потребуть чітких розрахунків або ж аналізу числових чи текстових даних. Йдеться не лише про видобування «патернів» чи трендів, а й також про семантичний аналіз – аналіз тексту на емоційність. [6]

- Удосконалений пошук. Розвиток NLP допомагає великим пошуковим системам точніше знаходити потрібний контент. Наприклад, система може краще розуміти запит, навіть якщо він сформульований в іншій граматичній формі, ніж текст у результатах пошуку. Крім того, тепер можна знайти матеріал не лише за його точною назвою, а й за коротким описом, який користувач сформулював своїми словами.

- Генерування контенту. Тексти, на які раніше людина могла витрачати купу часу, тепер може створювати і комп’ютер. Особливо це є корисним представникам креативних професій (копірайтери, комунікаційники, маркетологи, письменники), адже комп’ютер може враховувати побажання користувача щодо тону чи стилю листа, проте подібний інструмент стане у пригоді й іншим, зокрема задля написання шаблону листа, короткого підпису у соцмережах. [6]

- Розпізнавання та відтворення голосу. Цей аспект включає в себе як перетворення усного мовлення, прийнятого на вхід, у письмовий формат, так і голосове відтворення тексту комп’ютером. Зокрема це використовується в аудіовідповідачах, різних голосових чатботах, генерування аудіо та музики, відтворення тексту голосом із урахуванням різних стилів, інтонацій тощо.

Задачі NLP можна класифікувати за трьома рівнями:

- Низькорівневі. Такі задачі спрямовані на базову обробку тексту та його підготовку до складнішого аналізу. Приклади таких задач:
    - токенізація – розбиття тексту на одиниці
    - нормалізація – приведення слова до нижнього регістру, видалення пунктуації
    - лематизація/стемінг – приведення слова до її початкової форми
    - POS tagging (Part-of-Speech tagging) – визначення частини мови, до якої відноситься те чи інше слово.

Цікаво, що окремі експерти наводять дещо іншу характеристику низькорівневих задач:

- позначення мовних одиниць або пар одиниць відповідно до їхньої граматичності
- позначення мовних одиниць за типом емоції чи настроєм, які вони виражають
- позначення пар мовних одиниць відповідно до їхньої семантичної схожості
- позначення пар мовних одиниць відповідно до їхнього логічного зв’язку [7]

- Середньорівневі. Такі задачі потребують складнішої обробки, розуміння структури речення, синтаксису та граматичних зв’язків. Приклади:
    - Синтаксичний аналіз (парсинг) – побудова дерева речення, розбір граматичної структури речення
    - Chunking/shallow parsing – виявлення груп слів у речення (група іменника, група дієслова і тд.)
    - Аналіз залежностей/побудова онтології – визначення зв’язків між словами у реченням. Різниця між синтаксичним аналізом і аналізом залежностей полягає у тому, що синтаксичний аналіз визначає загальну структуру речення чи тексту, а аналіз залежностей дає розуміння синтаксичних зв’язків окремих слів між собою.

- Високорівневі. Ці задачі працюють із розумінням та генеруванням тексту. Наприклад:
    - Машинний переклад – переклад комп’ютером однієї мови на іншу
    - Sentiment analysis (аналіз тональностей та емоцій) – визначення тональності та емоційного забарвлення тексту чи окремих фраз
    - Підсумовування тексту – надання короткого висновку чи довідки щодо вхідного тексту
    - Генерування тексту – створення вихідного тексту
    - Діалогові системи (чат-боти, голосові асистенти) – системи спілкування з користувачем, відповідь яких імітує відповідь справжньої людини (фактично усі діалогові системи «виросли» з поняття фатичного діалогу)

1.3. Історія та розвиток NLP

Фердинанд де Сосюр, швейцарський лінгвіст початку XX століття, сформував ідею мови як системи, де значення виникає через взаємозв’язки між словами, ідеями та звуками. Хоч він помер до публікації своїх праць, його колеги Альбер Сешеє та Шарль Баллі зібрали лекційні матеріали й опублікували «Курс загальної лінгвістики» – фундаментальний текст структуралізму. Згодом його підходи вплинули на розвиток комп’ютерних мов та технологій. [8]

Дослідження в галузі NLP розпочалися наприкінці 1940-х років із проєктів машинного перекладу. Одним із перших проєктів машинного перекладу був проєкт Вівера і Бута (1946), що були натхенні досвідом криптоаналізу під час Другої світової війни. Особливої уваги ця тема набула після публікації «меморандуму Вівера» (1949), що запропонував ідеї використання криптографії та інформаційної теорії для перекладу.

Перші системи машинного перекладу мали обмеження: вони спрощено розглядати мови як набори слів і фраз із фіксованим порядком, ігноруючи багатозначність і глибшу семантику. Це призвело до низької якості перекладу та розчарування в технології. Новий поштовх дослідженням дала робота Ноама Хомського «Syntactic Structures» (1957), яка ввела поняття генеративної граматики. [1]

Паралельно з дослідженнями машинного перекладу, у 1950 році Алан Тьюрінг запропонував тест, щоб визначити, чи машина здатна «думати». Якщо машина може підтримувати розмову імітуючи людину так майстерно, що її не можна відрізнити від справжньої людини, то така машина вважається мислячою. У 1952 році модель Ходжкіна-Хакслі пояснила, як нейрони формують електричну мережу мозку. Ці дві події стали основою для розвитку штучного інтелекту та обробки природньої мови. [8]

У 1950-60-х роках у галузі NLP співіснували два підходи: лінгвістичний і статистичний. Повертаючись до ідеї машинного перекладу, дослідники вірили, що автоматичний якісний переклад стане можливим уже за кілька років, однак звіт ALPAC (1966) констатував обмежені результати та призвів до скорочення фінансування.

У 1970-х дослідження зосередилася на семантиці, репрезентації знань та діалогах. З’явились нові підходи: граматика ролей, семантичні мережі, концептуальні залежності. Створювалися й ранні прикладі системи, як-от ELIZA, SHRDLU, LUNAR, які демонстрували можливості комп’ютерного розуміння мови.

У другій половині 1970-х і 1980-х активно розвивались теорії дискурсу, комунікативних намірів та генерації природної мови. Водночас зросла зацікавленість у статистичних підходах як доповненні до символічних.

У 1990-х розвиток NLP прискорився завдяки збільшенню обсягів цифрового тексту, обчислювальної потужності й поширенню Інтернету. Статистичні методи стали домінуючими, зокрема для задач розпізнавання частин мови, розмежування значень слів тощо. [1] Тоді ж почались створюватися великі текстові корпуси, такі як Penn Treebank.

У 2000-х роках NLP став більш розвиненим та почав активно використовуватись у перекладачах, пошукових системах та голосових помічниках. Стали поширеними алгоритми SVM (Supported Vector Machines) та HMM (Hidden Markov Models). У 2010-х з'явилось глибинне навчання, яке суттєво просунуло розвиток NLP. У 2013 Google розробив Word2Vec – метод, що представляє слова у вигляді векторів та дозволяє машині розуміти семантичну схожість між словами. У 2017 році було представлено архітектуру Transformer, яка лягла в основу передових мовних моделей. OpenAI створив GPT (2018), а Google – BERT (2018), який розуміє слова у реченні незалежно від того, з якого кінця це речення буде прочитане машиною. Це стало проривом у розумінні мови комп'ютером.

У 2020-х NLP пережила нову еру з появою у світ GPT-3 (OpenAI) – моделі з 175 мільярдами параметрів, здатної генерувати тексти, відповідати на запитання та створювати контент на основі коротких підказок. Такі моделі відкрили можливості для автоматизації задач у сфері обслуговування клієнтів, створенні контенту, перекладі тощо. Також зростає увага до ефективності моделей, багатомовності, етичності та прозорості. [9] Сучасні системи NLP добре справляються з довільними текстами і враховують значну частину варіативності й неоднозначності природної мови. [1]

У майбутньому мовні моделі краще розумітимуть контекст запиту, що допоможе точніше інтерпретовувати людську мову. Покращення технологій багатомовної обробки тексту сприятиме зменшенню розриву між англійською та іншими мовами, сприяючи глобальній доступності. Завдяки мультимодальному навчанню (текст + зображення чи звук) моделі глибше розумітимуть складні поняття. Також розроблятимуться нові архітектури, здатними більш ефективно працювати з великими текстами. [10]

1.4. Складові NLP

1.4.1. Фонетичний та фонологічний аналіз

Фонетика – наука, що вивчає звуки мовлення: як вони утворюються, передаються та сприймаються. Фонетика вивчає алофони – реальні варіації вимови людиною фонем. Фонетика ділиться на три основі напрями:

- Акустична фонетика досліджує звукові хвилі, що утворюються органами мовлення й передаються через повітря до вуха слухача
- Аудиторна фонетика аналізує, як слуховий апарат і мозок сприймають звуки, зокрема досліджує фізіологічні процеси під час розпізнавання мовлення
- Артикуляційна фонетика вивчає, як рухаються органи мовлення (язик, губи, піднебіння тощо) під час вимови

Вирішення фонетичних проблем у природній мові передбачає роботу з омофонами – словами, що звучать однаково, але мають різне написання та значення. Вивчення омофон збагачує словниковий запас і допомагає краще розуміти значення слів у контексті. Фонетична ідентифікація тексту комп’ютером має такі кроки:

1. Створення словника омонімів, який міститиме як значення кожного слова, так і відповідні фонеми. Це допоможе розуміти, як звуки мають оброблятись та зберігатись
2. Зберігання звуків для кожного слова разом із його значенням. Це є складним завданням, зокрема через те, що машини не можуть сприймати мову у такий спосіб, як це роблять люди.
3. Застосування фонетичних алгоритмів, які присвоюють індекси звукам на основі їхньої вимови. Більшість таких алгоритмів розроблені лише для англійської мови, а тому не завжди є ефективними для інших мов.

Фонологія – це наука, яка зосереджена на фонемах – мінімальних одиницях звуку, що відрізняють значення слів у мові. Вона також описує правила поєднання звуків. [11] Фонологічний аналіз стосується аналізу звуків мовлення як усередині слів, так і між ними. Існує три типи правил у фонології:

- Фонетичні – описують звуки в межах слова
- Фонемні – пояснюють зміни вимови при поєднанні слів
- Просодичні – регулюють наголос і інтонацію в реченні

У системах NLP, що сприймають на вхід усне мовлення, звукові хвилі перетворюються в цифровий сигнал, який далі інтерпретується за цими правилами або шляхом порівняння з тією мовною моделлю, яка використовується. [1]

1.4.2. Морфологічний аналіз

Морфологічний аналіз стосується морфемної структури слів, тобто найменших одиниць значення. Наприклад, слово «узлісся» складається з чотирьох морфем: уз- (префікс), ліс (корінь), -с- (суфікс), я (закінчення). Задача комп'ютера в морфологічному аналізі така ж, як і у людини: розпізнати значення слова, розклавши його на частини. До того ж комп'ютер має також розпізнавати різні ознаки, якими наділені ті чи інші морфеми. Наприклад, суфікс -ed у дієсловах англійської мови вказує на минулий час – а ця смислова ознака може бути єдиним маркером часу в усьому реченні. [1]

Існує кілька задач морфологічного аналізу:

- Морфологічна сегментація – це задача обробки природньої мови, яка полягає в розділенні слів на окремі морфеми. Це важливий етап, оскільки допомагає зменшити проблеми, пов'язані з невідомими слоавми та нестачею даних, особливо в мовах із багатою морфологією.

- Лематизація – це процес приведення слова до його словникової форми (леми). Наприклад, слово «talked» перетворюється на «talk». Лематизація використовує лематизатори – спеціальні інструменти, що враховують граматичні правила для правильного зведення слова до базової форми.POS-розмітка (Part-of-Speech tagging) – це призначення кожному слову частини мови (наприклад, іменник, дієслово, прикметник тощо), а також додаткових граматичних ознак (множина, час, рід). Це необхідно для розуміння синтаксичної структури речення. POS-розмітку можна виконувати двома основними способами:
    - Правилами, які прописуються вручну
    - На основі корпусу – коли алгоритм навчається на великому наборі слів із вже розміченими словами

- Стемінг – це процес скорочення слови до основи (стему), яка до того ж не завжди є граматично правильною формою. Наприклад, fishing може скоротитися до fish. На відміну від лематизації, стемінг не враховує

морфологічних правил, а просто обрізає типові закінчення. Це робить його швидким, але менш точним. Іноді його вважають спрощеною формою лематизації. [12]

Для морфологічного аналізу програмістами були розроблені окремі бібліотеки мов програмування, зокрема, SpaCy, Stanza, NLTK, UDPipe, Morfessor, Foma/HFST, Polyglot, TreeTagger. Варіант реалізації морфологічного аналізу за допомогою бібліотеки SpaCy можна переглянути на ресурсі Kaggle: https://www.kaggle.com/code/samueljoseph502/morphological-analysis-nlp

1.4.3. Лексичний аналіз

Лексичний аналіз – це перший етап у багатьох NLP-процесах. Його мета – зрозуміти значення слів, їхній контекст і взаємозв'язки між словами. [13] На лексичному рівні як люди, так і NLP-системи інтерпретують значення окремих слів. Перше, що відбувається, - це визначення частини мови для кожного слова залежно від контексту.

Слова з єдиним значенням можуть бути одразу замінені на їхнє семантичне подання. Наприклад, слово launch (у значенні «великий човен») може бути подано через логічні предикати, що описують його клас, розмір і призначення:

launch (a large boat used for carrying people on rivers, lakes harbors, etc.)
((CLASS BOAT) (PROPERTIES (LARGE)
(PURPOSE (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))

Таке подання з використанням базових семантичних одиниць дозволяє уніфікувати значення слів і створювати складніші інтерпретації – подібно до того, як це робить людина

У лексичному аналізі часто використовується лексикон – словник, який може бути простим (лише слова і частини мови) або детальним, з інформацією про семантичні класи, аргументи, обмеження, значення та контексти використання багатозначних слів. [1]

Залежно від задачі, аналіз може мати різні форми: у компіляторах, наприклад, він перетворює код на послідовність «токенів» (окремі одиниці

тексту – слово, розділовий знак, числа, символи), ігноруючи пробіли та коментарі. У NLP групи слів можуть зберігатись як «n-грам» (послідовність із n елементів, зазвичай слів або символів, які йдуть підряд у тексті), зокрема це стосується стійких виразів. N-грами використовуються для аналізу частоти словосполучень, для побудови мовних моделей (наприклад, в автодоповненні) та врахування контексту в машинному перекладі, чат-ботах тощо.

Після токенізації система звертається до словника для визначення значень слів. У чатботах, наприклад, це може бути пошук інтенції користувача в базі даних. Через багатозначність слів, система повинна визначити правильне значення залежно від контексту. Щоб це зробити, слова в словнику часто пов’язують з типами контексту. Наприклад, «baseball field” може бути позначено як LOCATION для подальшого синтаксичного аналізу. [13]

1.4.4. Синтаксичний аналіз

Синтаксичний аналіз обробляє слова в реченні, щоб виявити його граматичну структуру. [1] Для цього потрібні граматика та парсер. Парсер – це програма або алгоритм, що аналізує структуру речення, згідно з певною граматикою, і розпізнає граматичні залежності між словами. Граматика – це набір правил, які визначають, чи можуть слова комбінуватись у реченні, щоб вони мали правильну структуру. Граматика допомагає парсеру зрозуміти, які елементи речення є підметом, присудком, додатком тощо.

Результат синтаксичного аналізу – структура, яка показує залежності між словами. Повний синтаксичний аналіз не завжди потрібен – іноді достатньо встановити зв’язки між фразами або частинами речення. Порядок слів і синтаксичні залежності важливі, бо впливають на значення. Наприклад:

«The dog chased the cat» (Собака погнався за котом)
«The cat chased the dog» (Кіт погнався за собакою)

відрізняються лише порядком слів, але мають інше значення. [1] Зрозуміло, що порядок слів є важливим далеко не у всіх мовах.
У синтаксичному аналізі є три основі види підходів:

- Правила-орієнтовані підходи

  - Контекстно-вільні граматики (CFG). Це традиційний метод, що використовує правила для опису поєднання між собою компонентів речення. Вони створюють дерево розбору, яке відображає синтаксичну структуру речення
  - Залежні граматики. Замість ієрархічних структур ці граматики використовують спрямовані зв’язки між словами, що показують, які слова залежать від інших. Особливо корисні для мов із вільним порядком слів.
- Статистичні підходи
  - Ймовірнісні контекстно-вільні граматики (PCFG). Це розширення контекстно-вільних граматик, яке додає до правил ймовірність. Це дозволяє визначити найімовірнішу синтаксичну структуру в контректному контексті.
  - Парсери на основі переходів. Використовують машинне навчання для поступової побудови дерева розбору, приймаючи рішення на кожному кроці.
- Підходи на основні нейронних мереж
  - Рекурентні нейронні мережі (RNN). Використовуються для задач синтаксичного аналізу, зберігаючи контекстну інформацію в прихованому стані. Однак такі мережі мають проблеми з обробкою довгих залежностей.
  - Згорткові нейронні мережі (CNN). Підходять для таких задач як залежний парсинг, оскільки добре захоплюють локальні шаблони між сусідніми словами
  - Моделі трансформерів. Моделі, як BERT чи GPT, захоплюють як локальні, так і глобальні синтаксичні залежності, показуючи найкращі результати у синтаксичному аналізі [14]

Кожен із цих підходів має свої переваги та недоліки, а вибір конкретного методу залежить від конкретної задачі. Правила-орієнтовані підходи більш зрозумілі, але менш гнучці, у той час як статистичні та нейронні добре справляються зі складними «патернами», проте потребують великих даних для навчання.

Головною технікою синтаксичного аналізу є парсинг. Парсинг – це процес розбору речення на граматичні компоненти та представлення їх у структурованій формі, наприклад, у вигляді синтаксичного дерева або графа залежностей. Основі алгоритки парсингу в NLP:

- Парсинг згори вниз (Top-Down Parsing) – це метод, який починається з найвищого рівня синтаксичної структури та рекурсивно розбиває її на дрібніші частинки. Він починається з верхнього правила граматики (наприклад, речення) і поступово застосовує нижчі правила для досягнення термінальних символів (слів). Якщо правило не застосовується, парсер «відкотиться» та спробує інші варіанти.

- Парсинг знизу вгору (Bottom-Up Parsing) – це підхід, який починається з окремих слів і поступово будує дерево розбору, комбінуючи слова в більші одиниці.

- Чарт-парсинг (Chart Parsing) – цей метод використовує динамічне програмування для створення структури даних, яка зберігає часткові дерева розбору. Він використовує алгоритм Earley або CYK для контекстно-вільних граматик. Чарт-парсери можуть пропонувати кілька варіантів розбору для складних синтаксичних структур

- Shift-reduce парсинг – приклад стратегії згизу вгору, де слова додаються до стеку, а потім зменшуються за допомогою граматичних правил. Цей ефективний метод може обробляти непроективні синтаксичні структури. [14]

1.4.5. Семантичний аналіз

Семантичний аналіз – це метод обробки природної мови, який передбачає вивчення значення слів і фраз для розуміння автора в реченні чи абзаці. Він також вивчає контекстуальні зв’язки між елементами мови, що дозволяє глибше зрозуміти текст. Це зазвичай досягається шляхом виявлення ключових ідей і зв’язків у тексті за допомогою алгоритмів і штучного інтелекту. [15]

Цей рівень обробки часто вважається таким, який визначає значення слова, проте, як бачимо з вище наведеної інформації, значення формують усі рівні аналізу. Семантичний аналіз у свою чергу визначає можливі значення речення, зосереджуючись на взаємодії значень слів у ньому. Цей рівень включає семантичне розрізнення слів з кількома значеннями, подібно до синтаксичного розрізнення слів, що можуть виконувати різні граматичні ролі. Семантичне розрізнення дозволяє вибрати лише одне значення для багатозначних слів. [1]

Наприклад, слово «коса» може мати кілька значень: знаряддя праці для скошування трави, зачіска або ознака предмета – коса, тобто похила (прикметник). Для вирішення, яке ж значення слово має в тому чи іншому контексті, семантичний рівень аналізу використовує інформацію з решти речення, щоб вибрати правильне значення. Різні методи можуть використовувати частотність значень у корпусі, локальний контекст або теорії з прагматики. [1]

Семантичний аналіз використовує різні методи, але всі вони спрямовані на розуміння тексту так, як це робить людина. Два найпопулярніші методи семантичного аналізу такі:

- Поєднання машинного навчання та обробку природної мови для виявлення основних зв’язків та ідей у тексті. Це може включати використання моделі машинного навчання, навченої на великому обсязі текстів, виявлення їхніх ключових ідей та взаємозв’язків.

- Використання попередньо визначених онтологій та структурованих баз даних, що описують концепти і зв’язки в певній галузі. Це дозволяє алгоритмам швидше знаходити і витягати потрібну інформацію з тексту.

У семантичному аналізі також є важливою лексична семантика, адже вона дозволяє комп’ютерам розуміти зв’язки між лексичними елементами (словами, стійкими виразами, фразовими дієсловами):

- Гіпонімія відображає зв’язок між загальним терміном і його конкретними застосуваннями
- Омонімія – подібність між двома словами, коли вони мають однакову форму або написання, але мають різне значення

- Полісемія – подібність між двома словами, коли вони мають однакове написання і їхні значення схожі, проте не конкретні
- Синонімія – відношення між двома словами, різними за написанням, але схожим за значенням
- Антонімія – відношення між двома словами, різними за написанням та різними за значенням
- Меронімія – спосіб поєднання тексту або слів так, щоб вони мали сенс та вказували на одне поняття як складову іншого. [15]

1.4.6. Прагматичний аналіз

Прагматичний аналіз стосується цілеспрямованого використання мови в конкретних ситуаціях і враховує контекст, що виходить за межі самого тексту. Його мета – пояснити, як ми розуміємо додаткові значення, які не прямо виражені в тексті. Для цього потрібні знання про світ, включаючи розуміння намірів, планів та цілей. У деяких NLP-застосунках використовуються бази знань та модулі формулювання логічних висновків.

Наприклад, у реченнях:

«Міська рада відмовила демонстрантам у дозволі, бо вони боялися насильства»

«Міська рада відмовила демонстрантам у дозволі, бо вони закликали до революції»

щоб зрозуміти, хто саме мається на увазі під «вони», потрібні знання про типову поведінку та мотивацію учасників подій – тобто прагматичне чи світове знання. [1] Ключовим елементом розуміння є контекст, він є трьох видів:

- Дискурсний – де саме розміщене висловлювання у тексті
- Фізичний – враховує фізичні умови ситуації
- Соціальний – хто говорить, до кого і з якою метою [16]

1.4.7. Дискурсний аналіз

На відміну від синтаксису та семантики, які працюють із одиницями довжиною в одне речення, рівень дискурсу в обробці природної мови працює з текстами, що складаються з кількох речень. Дискурс не інтерпретує багатореченеві тексти як просто поєднані окремі речення. Замість цього він зосереджений на зв'язках між реченнями, які надають тексту значення.

До основних завдань дискурсу відносяться розв'язка анафори та визнання структури тексту. Розв'язка анафори замінює слова, такі як займенники, на відповідні сутності, до яких вони відносяться. Визнання структури тексту визначає функції речень у тексті, що додає до його змістовного представлення. Наприклад, газетні статті можуть бути розібрані на складові дискурсу, такі як: введення, основна історія, попередні події, оцінка, цитати та очікування. [1]

Інші завдання дискурсу включають:

- Генерація природної мови (NLG) – процес створення наративів або описів природною мовою з упорядкованих даних.
- Витяг інформації та резюмування тексту.
- Машинний переклад – процес перекладу тексту з однієї мови на іншу за допомогою обчислювальних моделей, здатних точно розуміти семантичне групування елементів.
- CALL (Комп'ютерна підтримка вивчення мов) – використання комп'ютерних технологій у навчанні мовам. [16]

1.5. Висновок до розділу 1

У цьому розділі курсової роботи розглянуто основи обробки природної мови (NLP) - підгалузі штучного інтелекту, що забезпечує можливість комп'ютерів розуміти та обробляти людську мову. Задачі NLP охоплюють численні аспекти, від базової обробки тексту до складних завдань, таких як генерація та розуміння мови, що дозволяє автоматизувати рутинні завдання, покращувати пошук інформації, здійснювати аналіз даних і генерувати контент.

NLP можна поділити на кілька основних напрямів: Natural Language Understanding (NLU), що спрямоване на розуміння змісту тексту, та Natural Language Generation (NLG), яке фокусується на створенні тексту. Прогрес у цій сфері значною мірою залежить від розвитку мовних моделей і технологій, таких як машинний переклад, аналіз емоцій, генерація тексту та діалогові системи.

Ця галузь значно впливає на різноманітні сфери життя, зокрема на фінанси, охорону здоров'я, юриспруденцію, і є невід’ємною частиною сучасних технологій. Водночас існують різні рівні складності задач NLP, від базових операцій з текстом до складного аналізу семантики та синтаксису.

Розвиток NLP тісно пов'язаний з етапами становлення штучного інтелекту, починаючи з перших спроб машинного перекладу після Другої світової війни до сучасних інновацій у розпізнаванні мовлення та генерації тексту. Сьогодні спеціалісти в галузі NLP є одними з найбільш затребуваних фахівців у технологічних компаніях, що підтверджує важливість цієї сфери для розвитку сучасних інформаційних технологій.

NLP складається з декількох рівнів обробки: фонетичний та фонологічний, морфологічний, лексичний, синтаксичний, семантичний, прагматичний та дискурсний аналіз.

РОЗДІЛ 2. СТАН NLP В УКРАЇНІ

2.1. Історичний огляд

2.1.1. Початок NLP в Україні

Про історичний розвиток NLP в Україні відомо мало. Відомо однак, що одним із піонерів галузі NLP в Україні був Тарас Вінцюк, який розробив свою широко застосовувану генеративну модель для розпізнавання образів у 1967 році, відому як Dynamic Time Warping (DTW). Цей підхід використовується в теорії розпізнавання мовлення та образів, а також у текстовому процесингу, радіофізиці та біології. Подібну модель, Hidden Markov Model (HMM), було створено в 1973 році, і вона стала найцитованішою у світі.

З кінця 1960-х років команди на чолі з Вінцюком розробляли системи розпізнавання мовлення. Він автор понад 300 наукових праць та двох книг. У 1988 та 1997 роках Вінцюк отримав Державну премію України в галузі науки та техніки.

Тарас Вінцюк також є творцем концепції "Комп'ютера для розпізнавання образів", що стала основою Національної науково-технічної програми (2000–2010). Він був членом багатьох наукових товариств, організував 10 міжнародних конференцій "UkrObraz" і започаткував щорічні літні школи з мовних технологій. [17]

Пізніше були засновані різноманітні інституції, орієнтовані на дослідження галузі штучного інтелекту, зокрема Київська лабораторія штучного інтелекту та Київський інститут проблем штучного інтелекту. В основні напрями діяльності другого, зокрема входила розробка розпізнання мовленнєвих та зорових образів, створення природномовних інтерфейсів та систем розпізнавання мовних образів. [18]

2.1.2. UNLP

Українська спільнота з обробки природної мови (NLP) почала формуватися лише нещодавно, здебільшого через ізольовані дослідження окремих груп. Щоб об'єднати ці зусилля, Український воркшоп з NLP (UNLP), започаткований Українським католицьким університетом та НаУКМА, став майданчиком для обміну ідеями та співпраці. Від того часу воркшопи проводяться щорічно:

UNLP 2021. Перший воркшоп відбувся у Херсоні у гібридному форматі, паралельно з конференцією ICTERI 2021. Ключовим спікером був Джон МакКрей з Ірландії.

UNLP 2023. Другий UNLP пройшов онлайн у межах EACL 2023. Попри війну, подія зібрала близько 100 учасників. Вперше проведено Shared Task зі створення систем для виправлення граматичних помилок українською.

UNLP 2024. Третій UNLP відбувся разом з конференцією LREC-COLING 2024 у гібридному форматі. Іван Вулич, Василь Старко та Андрій Рисін виступили з доповідями про багатомовні системи діалогу та ресурси для української. У межах воркшопу презентували 16 дослідницьких робіт. [19]

2.2. Огляд розроблених іструментів

2.2.1. Великий електронний словник української мови

ВЕСУМ (https://github.com/brown-uk/dict_uk) – це словник словозміни української мови, основними компонентами якого є реєстр лем, коди класів словозміни й правила генерації словоформ на основі цих кодів, а також із застосуванням елементів програмової логіки.

У ВЕСУМ-і морфологічний аналіз і синтез мають також динамічний характер. Це потрібно для обробки слів, які утворюються за типовими моделями, але не охоплені словником – зокрема складних прикметників, іменників, прислівників на "по-", слів із частинами типу "бізнес-" чи "онлайн-". Завдяки розбиттю таких слів на складники й зверненню до словника, система правильно визначає їхні граматичні ознаки у 95% випадків. Цей підхід називається «динамічним тегуванням».

ВЕСУМ має низку характеристик, зокрема:

- Машинозчитування – інтегрується з LibreOffice, Word, Google Docs, браузерами Firefox і Chrome, а також з Apache Lucene.
- Відкритість – можна вільно користуватись і долучатись до розвитку.
- Динамічність – постійне оновлення словника.
- Масштаб – понад 401 тис. лем, з яких генерується 6 млн словоформ (3,4 млн – унікальні).
- Власні назви – охоплює всі населені пункти України, декомунізовані назви, тисячі імен, прізвищ, по батькові, іноземні топоніми.
- Різноманітність лексики – включає покручі, абревіатури, сленг, неофіційні форми та рідковживані слова.
- 13 словозмінних класів – охоплюють традиційні частини мови та додаткові категорії (іншомовні, незмінювані тощо).
- Гнучкість – хоча наголоси не вказано, їх можна додати за потреби.
- Зручна система тегів і кодів – спрощує додавання нових слів.
- Відмінювання складних назв. [20]

та багато інших.

Зображення 2.2.1.1 Приклад використання ВЕСУМ-а

https://vesum.nlp.net.ua/?w=ліс

2.2.2 БРУК

Браунський корпус (Brown Corpus) – перший машиночитний корпус сучасної англійської мови, створений у 1960-х роках у Браунському університеті (США) В. Н. Френсісом і Г. Кучерою. Він містить 1 млн слів із 500 уривків редагованих текстів, опублікованих у 1961 році. Цей корпус став зразком для подальших корпусів: британського ЛОБ, американського Фраун і британського Ф-ЛОБ, що відображають мовну ситуацію початку 1990-х років. За його моделлю також створено корпуси для інших мов, зокрема болгарської.

Такі корпуси забезпечують збалансоване представлення мови, сприяють лінгвістичним і міжмовним дослідженням, є корисними для вивчення мови й корпусної лінгвістики, а також дозволяють тестувати методи корпусного аналізу. [21]

БрУК – відкритий, жанрово збалансований корпус сучасної української мови обсягом 1 млн слововживань, створений за принципами англійського Brown Corpus і призначений для подальшої анотації. [22]

Станом на 2024 рік БрУК налічував 600 тисяч слів зі знятою омонімією, а саме:

- 600130 українських токенів
- 746845 токенів загалом
- 608782 токенів-слів і чисел
- 601970 токенів слів, що складаються лише з букв
- 103781 унікальне українське слово
- 95534 унікальні українські слова (без урахування регістру)

- 44179 унікальних лем [23]

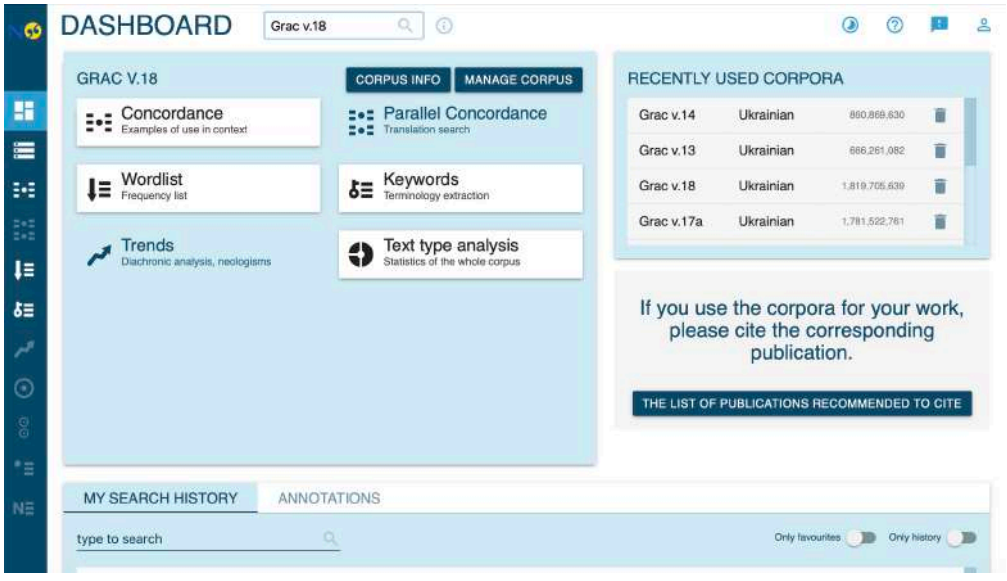
2.2.3 ГРАК

Генеральний регіонально анотований корпус української мови (ГРАК) –це великий репрезентативний корпус текстів українською мовою, що охоплює період від 1816 року до сьогодення та представлений у різних стилях, жанрах і регіонах. Корпус містить понад 150 тисяч текстів від близько 35 тисяч авторів. Кількість токенів становить понад 600 млн.

Через зручний інтерфейс користувачі можуть шукати слова, граматичні форми та їх поєднання, фільтрувати пошук за підкорпусами (час, регіон, стиль тощо), отримувати статистику й створювати вибірки. Регіональна анотація дає змогу враховувати походження тексту й автора, що особливо важливо з огляду на історичну варіативність української мови.

ГРАК є корисним інструментом для лінгвістичних досліджень, створення словників, підручників і навчальних матеріалів на основі реальних прикладів мови. [24]

До наповнення корпусу ГРАК регулярно долучаються студенти та викладачі різних українських університетів, зокрема Національного технічного інституту «Харківський політехнічний інститут», Одеський національний університет І.І. Мечникова, Львівського національного університету імені Івана Франка, Національного університету «Києво-Могилянська академія» та інших.



Зображення 2.2.3.1 Головна сторінка ГРАК-18

https://sketch.uacorpus.org/#dashboard?corpname=grac18

2.3.4 LanguageTool API NLP UK

LanguageTool API NLP UK — демонстраційний проєкт для обробки української мови за допомогою LanguageTool. Він реалізований на Groovy з допоміжними обгортками для Python3 та Java. Основні інструменти дозволяють токенізувати (TokenizeText.groovy) та тегувати (TagText.groovy) тексти, використовуючи формат UD (Universal Dependencies). Для роботи необхідно встановити JDK 17 і Groovy.

Функції:

- Токенізація текстів
- Морфологічний аналіз і тегування
- Виправлення типових помилок у тексті
- Генерація списку невідомих слів

Інструмент базується на українському модулі LanguageTool та словнику з проєкту ВЕСУМ. Проєкт стане в пригоді для:

- розробників, які працюють з українською мовою;
- дослідників у сфері лінгвістики;
- створення навчальних корпусів;
- інтеграції в текстові редактори чи онлайн-сервіси. [25]

2.3.5 Інші інструменти

Останні роки спостерігається активне зростання кількості відкритих ресурсів для обробки природної мови українською мовою, що дає змогу будувати високоякісні системи NLP (Natural Language Processing).

Корпуси та набори даних охоплюють як одномовні, так і паралельні ресурси. Одномовні корпуси включають Malyuk (113 ГБ текстів із UberText 2.0, OSCAR та новинних джерел), UberText 2.0 (більше 5 ГБ новин, Вікіпедії, художньої, соціальної та правової літератури), OSCAR, CC-100, mC4 – автоматично зібрані та очищені корпуси з Common Crawl, які містять десятки гігабайтів українських текстів. Також доступні спеціалізовані набори, зокрема корпус українських твітів для виявлення токсичності, 250 тисяч речень із форумів, а також понад 3 мільйони заголовків новин.

Паралельні корпуси включають ресурси від OPUS, Tatoeba, польсько-український корпус, а також синтетичні переклади з Вікіпедії та корпус Wiki Edits з понад 5 мільйонів пар речень. Для задач із анотацією також доступні спеціалізовані ресурси: корпус ЗНО (~4000 питань), UA-GEC для виправлення граматичних помилок, UA-SQuAD (українська версія Stanford QA), корпуси для розпізнавання іменованих сутностей (NER-uk) та кореференції (Ukrainian OntoNotes). Також використовується Universal Dependencies — корпус синтаксичних дерев залежностей.

Окрему роль відіграють словники та лінгвістичні ресурси: ВЕСУМ (словник частин мови) – про який згадувалось вище, словник наголосів (2.7 млн форм), тональний словник, словник ненормативної лексики (obscene-ukr), а також список абревіатур.

У сфері інструментів обробки тексту доступні як класичні, так і сучасні бібліотеки: pymorphy2 (лематизатор і POS-аналізатор), LanguageTool (перевірка граматики і стилю), Stanza, NLP-Cube та nlp-uk – з підтримкою токенізації, морфологічного аналізу, синтаксису та розпізнавання сутностей.

Розроблено й адаптовано чимало попередньо натренованих моделей. До авторегресивних належать: aya-101 (13B параметрів), pythia-uk (на основі mT5), UAlpaca (LLaMA, натренована на інструкціях), GPT-2-подібні моделі (Tereveni-AI, uk4b). Серед маскованих моделей – xlm-roberta-base-uk, youscan/ukr-roberta-base. Для перекладу активно використовуються моделі Helsinki-NLP / OPUS-MT та M2M-100, що підтримують десятки мов. Послідовні моделі типу mBART50 і mT5 дають змогу вирішувати широкий спектр завдань: від переформулювання до узагальнення.

Для частиномовного аналізу доступні моделі lang-uk/flair-uk-pos, а для розпізнавання іменованих сутностей – ukr-models/uk-ner, flair-uk-ner, MITIE. У задачах, пов’язаних із векторним представленням слів, застосовуються fastText (тренований на Wikipedia і CommonCrawl), Word2Vec, GloVe, BPEmb, а також українські варіанти моделей для підрахунку співзвучності (Flair, LexVec).

Окрему категорію становлять моделі для відновлення пунктуації та регістру, зокрема uk-punctcase та punctuation_uk_bert, а також інструменти для встановлення наголосів (ukrainian-word-stress).

Також існують комерційні набори даних, як-от LORELEI Ukrainian Representative Language Pack, і платформи для розпізнавання та синтезу мовлення, наприклад, egorsmkv/speech-recognition-uk.

Активно проводяться дослідницькі події, як-от Ukrainian NLP Workshop (UNLP) – про який згадувалось вище – який включає щорічні змагання: GEC (2023), LLM-файнтюнінг (2024) і виявлення маніпуляцій у соцмережах (2025). Ці заходи сприяють створенню відкритих датасетів, інструментів і підвищують якість українських NLP-систем. [26]

2.3 Висновок до розділу 2

Історія обробки природної мови (NLP) в Україні бере початок ще з 1960-х років, коли Тарас Вінцюк розробив генеративну модель для розпізнавання образів — Dynamic Time Warping (DTW), яка й сьогодні застосовується в мовленнєвих і текстових технологіях. Під його керівництвом створювалися системи розпізнавання мовлення, а сам Вінцюк зробив значний внесок у розвиток української науки, започаткувавши концепцію "комп’ютера для розпізнавання образів" і організовуючи конференції "UkrObraz" та літні школи з мовних технологій.

Подальший розвиток NLP в Україні здійснювався через створення наукових інституцій, зокрема Київського інституту проблем штучного інтелекту, що спеціалізувався на розпізнаванні мовленнєвих образів і створенні природномовних інтерфейсів.

З 2021 року формується сучасна українська NLP-спільнота через серію щорічних заходів UNLP, організованих Українським католицьким університетом і НаУКМА. Перший воркшоп відбувся у Херсоні, другий — онлайн у межах EACL 2023 (із запуском Shared Task для виправлення граматичних помилок українською), третій – у межах LREC-COLING 2024 з 16 дослідницькими доповідями.

Останнім часом активно розвиваються різні відкриті ресурси для обробки української мови. До таких інструментів належать:

- Корпуси: Malyuk, UberText, OSCAR та інші, що містять десятки гігабайтів українських текстів.
- Паралельні корпуси: OPUS, Tatoeba, синтетичні переклади з Вікіпедії, для полегшення перекладу та анотацій.
- Спеціалізовані ресурси: Корпуси для виявлення токсичності в українських твіттах, а також ресурси для виправлення граматичних помилок (UA-GEC) і відповіді на запитання (UA-SQuAD).
- Лінгвістичні ресурси: ВЕСУМ, словники наголосів, ненормативної лексики, абревіатур, а також лексичні моделі на базі fastText, Word2Vec, GloVe.
- Моделі та бібліотеки: Моделі для переформулювання, узагальнення (mT5, mBART), для частиномовного аналізу (flair-uk-pos) і розпізнавання іменованих сутностей (ukr-models/uk-ner).

Ці ресурси та інструменти активно використовуються в NLP-системах для обробки української мови, зокрема для автоматичного перекладу, аналізу тексту та створення нових лінгвістичних моделей.

РОЗДІЛ 3. ПРОБЛЕМИ В УКРАЇНСЬКОМУ NLP

3.1 Мовні особливості

Українська мова має низку лінгвістичних характеристик, які суттєво впливають на обробку природної мови (NLP). Ці особливості зумовлюють потребу в спеціалізованих моделях, корпусах та алгоритмах, адаптованих саме до української.

1. Флективність та багата морфологія

Українська є флективною мовою, що означає велику кількість змінних форм одного слова залежно від граматичних категорій (відмінок, рід, число, час, аспект тощо). Наприклад, іменник «рука» має 14 форм (7 відмінків у двох числах), а дієслова – ще більше.

Це ускладнює задачі токенізації, лематизації, частиномовного аналізу та машинного перекладу, оскільки система має враховувати багату парадигму словозміни.

2. Вільний порядок слів

У типових слов'янських мовах, зокрема українській, відносно вільний порядок слів у реченні (SVO, OVS, VSO та інші комбінації) дозволяє змінювати структуру без втрати змісту. Це створює виклик для моделей синтаксичного аналізу (dependency parsing), які в англомовних моделях часто ґрунтуються на фіксованому порядку слів.

3. Фонетичні особливості та наголос

Наголос в українській мові є рухомим і смислорозрізнювальним (напр., за́мок – будівля, замо́к – пристрій). Більшість текстів не мають маркера наголосу, що ускладнює завдання TTS (text-to-speech), автоматичного наголошування та роботи з поезією. Потрібні окремі словники та моделі наголошування, як-от ukrainian-word-stress або словник наголосів (2.7 млн форм, створений на основі ВЕСУМ).

4. Синонімія, варіативність та стилістичне багатство

Українська мова має розвинену систему синонімів, а також паралельне використання книжних і розмовних форм (напр., багато vs купу, розпочати vs почати). Це важливо для задач семантичного аналізу, аналізу сентименту, стилістичного переформулювання та генерації тексту. Лексична варіативність ускладнює задачу класифікації наміру, пошуку релевантної інформації та машинного перекладу.

5. Складна система префіксів і суфіксів

Українська має продуктивну словотвірну систему: один корінь може створити десятки форм за допомогою різних афіксів (писати, написати, записати, переписати, дописати, тощо). Це вимагає врахування словотвірних ланцюгів у задачах лематизації, розпізнавання нових слів, та у трансформерах – правильного токенізування (наприклад, byte-pair encoding).

6. Кальки та запозичення в умовах війни та диджиталізації

Сучасна українська мова інтенсивно запозичує слова з англійської (зокрема в IT, медіа, військовій лексиці): дрон, тейкдаун, скан, донат, реквест. Часто такі слова вживаються у зміненій формі або створюються кальки (напрямок політики ← policy direction). Це створює виклики для токенізації, морфологічного аналізу, аналізу сентименту та NER.

7. Діалекти та регіональні варіанти

Попри стандартизацію, українська має чітко виражені діалекти: південно-західний, північно-східний, середньонаддніпрянський та вплив суржику в окремих регіонах. У задачах класифікації, модерації в соцмережах, автоматичного перекладу або генерації такі варіанти мови потребують окремого врахування або фільтрації.

3.2 Основні проблеми

У межах корпусної лінгвістики української мови виділяють кілька ключових проблем:

1. Розмітка тексту. Це включає токенізацію (поділ тексту на окремі слова), лематизацію (зведення словоформ до початкової форми) та морфологічний аналіз. Важливість цих процесів зумовлена складною флективною системою української мови, яка потребує точної обробки.

2. Репрезентативність корпусу. Для відображення типових лексико-граматичних явищ потрібна велика кількість мовного матеріалу. Наприклад, для достовірного аналізу перших 5 тисяч найуживаніших слів необхідно принаймні 10–20 млн слововживань, а для перших 20 тисяч – понад 100 млн, що зумовлено Законом Ципфа.

3. Подання результатів. Великі корпуси породжують масивні вибірки результатів запитів, тому необхідне автоматичне групування (кластеризація) знайдених прикладів та побудова статистичних моделей колокацій (стійких словосполук).

4. Веб як джерело корпусу. Інтернет-матеріали можна використовувати як корпус, однак вони здебільшого не містять лінгвістичної розмітки (наголоси, граматичні категорії тощо). Хоча автоматизоване збирання сторінок і наступна розмітка дозволяють оперативно формувати корпуси для менш досліджених мов, такі тексти відображають більше комунікативні потреби користувачів, ніж систематичні мовні закономірності. [27]

Українські розробники також висловлюються про певні проблеми обробки природної мови, що стримують розвиток галузі:

- Машини погано справляються з багатозначністю, синонімією, омонімією, метафорами, грою слів і контекстом. Постає виклик: навчити алгоритми "розуміти", що саме має на увазі людина – не просто слова, а сенс, емоції й ситуацію.
  Приклад: речення «графік (художник) не вписався в графік (план)» складно правильно інтерпретувати машині без контексту.

- Існує проблема визначення емоцій і тональності, складність у виявленні емоційного тону повідомлення, особливо якщо там є сарказм, іронія чи культурні особливості. Машина має зрозуміти, що "позитивне" чи "негативне" – це не лише окремі слова, а контекст і культурна інтерпретація.
  Приклад: фраза «готель гарний, але ресторан не сподобався» — змішана тональність, яку складно оцінити автоматично.

- Кросмовна морфологія і малоресурсні мови. Більшість мов світу не мають достатньої кількості цифрових ресурсів для повноцінної роботи з ними в NLP. Внаслідок цього – обробка мов із незначною кількістю даних, відсутністю письма, словників або корпусів текстів.

- Проблема гумору і креативності. Машинам важко створювати або розпізнавати гумор, іронію, натяки, культурні алюзії, які є важливою частиною живої мови. Варто не лише навчити їх просто розуміти прямий

зміст, а й жартувати, бути дотепним, «читати між рядків». До прикладу у Стенфорді розробили алгоритм Хьо Хьо, який навчається створювати каламбури. Подібні технології – лише перший крок у цьому напрямку. [28]

3.3 Можливі шляхи подолання проблем

У контексті викликів, що постають перед природною обробкою української мови, існує низка перспективних напрямів, здатних суттєво покращити якість обробки та аналізу мовного матеріалу:

Удосконалення інструментів автоматичної розмітки. Розробка точніших алгоритмів токенізації, лематизації та морфологічного аналізу з урахуванням флективної природи української мови має стати пріоритетом. Зокрема, слід застосовувати методи гібридного моделювання – поєднання правил на основі граматики з машинним навчанням. Перспективними є також трансформерні архітектури (на кшталт BERT, mBERT, XLM-R), адаптовані під українську мову.

- Наповнення поточних корпусів більшою кількістю даних. Розширення обсягів корпусів, зокрема через краудсорсинг, співпрацю з видавництвами, науковими установами та медіа. Для покращення репрезентативності варто дотримуватись принципу збалансованості щодо жанрів, регіональних і соціальних варіацій. Також можливе застосування інструментів автоматичного добору корпусів з вебу (web crawling), доповненого розміткою на основі машинного навчання.

- Візуалізація та інтерпретація даних. Для ефективного використання корпусів необхідні зручні інтерфейси для пошуку та аналізу, що дозволяють кластераналіз результатів, побудову статистичних моделей колокацій, візуалізацію мереж лексичних зв’язків. Наприклад, застосування бібліотек для побудови графів (GraphViz, Gephi) чи інтерактивних дашбордів на базі Python/R.

- Подолання багатозначності й контекстної неоднозначності. Використання контекстно-залежних мовних моделей (contextual embeddings) здатне підвищити точність інтерпретації значень, омонімії та синонімії. Впровадження моделей типу ELMo, BERT, GPT, які

враховують контекст, відкриває нові можливості в розв’язанні проблем багатозначності.

- Розвиток емоційного та прагматичного аналізу. Для кращого визначення тональності повідомлень слід навчати моделі на локалізованих наборах даних, які враховують культурні особливості, іронію та сарказм. Створення вручну анотованих корпусів для тренування моделей тональності з врахуванням змішаних оцінок – актуальне завдання для україномовного NLP.

- Підтримка малоресурсних мов та діалектів. Одним зі способів підтримки є перенесення знань із високоресурсних мов (transfer learning), використання спільного мультилингвального простору та навчання моделей перекладу між близькими мовами (наприклад, між українською, білоруською та польською). Також варто підтримувати ініціативи відкритого доступу до корпусів, словників і мовних моделей.

- Розпізнавання гумору та креативності. Для аналізу креативних мовних явищ необхідне залучення корпусів, що містять приклади жартів, мемів, каламбурів. Перспективним є поєднання лінгвістичного та когнітивного аналізу, а також створення спеціалізованих підзадач для тренування моделей (наприклад, розпізнавання іронії в соціальних мережах).

Таким чином, подолання зазначених проблем передбачає міждисциплінарну співпрацю між лінгвістами, програмістами, філологами та представниками сфери штучного інтелекту, а також активну підтримку з боку держави й академічної спільноти. За таких умов корпусна лінгвістика української мови зможе подолати поточні виклики та розвиватись у новому темпі.

3.4 Висновок до розділу 3

Обробка природної мови української мови (NLP) стикається з рядом специфічних лінгвістичних та технічних викликів, зумовлених її флективною природою, вільним порядком слів, багатою морфологією та лексичною варіативністю. Окрім цього, на розвиток NLP в Україні впливає активне використання англіцизмів, діалектів та специфічних термінів, пов'язаних із сучасними соціальними та політичними процесами, зокрема в умовах війни.

Проблеми, які виникають у межах корпусної лінгвістики та галузі обробки природної мови, включають складнощі в розмітці тексту, обмежену репрезентативність корпусів, а також важливість забезпечення точності результатів через автоматичне групування та кластеризацію великих обсягів даних. Крім того, проблема багатозначності, контекстної неоднозначності та гумору, а також емоційного аналізу текстів, залишається однією з основних для розвитку NLP.

Для ефективного вирішення цих проблем необхідне удосконалення інструментів автоматичної розмітки та морфологічного аналізу, наповнення корпусів більшою кількістю даних, а також інтеграція новітніх мовних моделей, таких як трансформери (BERT, mBERT), які здатні враховувати контекст і багатозначність. Окрім цього, підтримка малоресурсних мов та діалектів, розвиток емоційного аналізу та розпізнавання гумору є важливими напрямками для досягнення високої точності та адекватності в обробці україномовного тексту.

Висвітлені шляхи подолання цих проблем вимагатимуть міждисциплінарної співпраці та активної підтримки з боку державних і академічних структур, що дозволить створити необхідні ресурси та інструменти для подальшого розвитку української NLP. У результаті це відкриє нові можливості для ефективного використання сучасних технологій штучного інтелекту в різних сферах, від машинного перекладу до створення інтелектуальних систем, що працюють з українською мовою.

ВИСНОВКИ

У курсовій роботі було розглянуто ключові аспекти обробки природної мови (NLP) з огляду на її теоретичні основи, історичний розвиток в Україні та специфічні проблеми, що виникають при обробці української мови. У першому розділі підкреслено важливість NLP як підгалузі штучного інтелекту, яка дозволяє комп'ютерам взаємодіяти з людською мовою, що має широкий спектр застосувань від автоматизації рутинних завдань до аналізу даних і генерації тексту. Розвиток цієї технології значно вплинув на різні сфери, зокрема фінанси та охорону здоров'я, та продовжує відігравати важливу роль у сучасних інноваціях.

Другий розділ присвячений історії розвитку NLP в Україні, де було відзначено перші досягнення в цій галузі, зокрема роботу Тараса Вінцюка та розвиток наукових інституцій, що сприяли розпізнаванню мовлення та створенню природномовних інтерфейсів. Сучасна українська NLP-спільнота активно розвивається завдяки інструментам і ресурсам, таким як корпуси та лінгвістичні моделі, що сприяють розвитку технологій для автоматичного перекладу та аналізу тексту.

Третій розділ зосереджений на специфічних викликах, з якими стикається NLP при обробці української мови. Зокрема, складнощі виникають через багатозначність, контекстну неоднозначність, вільний порядок слів та багату морфологію. Для подолання цих проблем важливим є вдосконалення інструментів автоматичної розмітки, збільшення репрезентативності корпусів, а також інтеграція новітніх мовних моделей. Розвиток цих технологій потребує міждисциплінарної співпраці та підтримки з боку державних і академічних структур.

Таким чином, дослідження підтверджує, що NLP є важливою складовою розвитку штучного інтелекту, і для подальшого успіху в обробці української мови необхідно активно працювати над подоланням існуючих лінгвістичних та технічних проблем, що дозволить досягти високої точності й ефективності в розпізнаванні та генерації україномовного контенту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub
2. “natural language”. Cambridge Dictionary | English Dictionary, Translations & Thesaurus. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://dictionary.cambridge.org/dictionary/english/natural-language#google_vignette
3. A. Chopra, A. Prashar та C. Sain, “Natural Language Processing”, *INT. J. TECHNOL. ENHANCEMENTS EMERG. ENG. RES.*, т. 1, № 4, б. д. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eeace1d14e266a5cd44fe781a874c662928602fd
4. “Timothee Mickus”. Home – Timothee Mickus. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://timotheemickus.github.io/my-phd-explained-to-my-folks/2021/07/18/computational-linguistics-vs-nlp.html
5. E. Kavlakoglu та R. Vaish. “NLP vs. NLU vs. NLG: What's the Difference? | IBM”. IBM - United States. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.ibm.com/think/topics/nlp-vs-nlu-vs-nlg
6. IBM. “What Is NLP (Natural Language Processing)? | IBM”. IBM - United States. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.ibm.com/think/topics/natural-language-processing
7. S. McRoy та C. Ali, *Principles of Natural Language Processing*. McRoy, SusAn, 2021. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://wisconsin.pressbooks.pub/naturallanguage/chapter/benchmarktasks/
8. “A Brief History of Natural Language Processing - DATAVERSITY”. DATAVERSITY. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/
9. “Master NLP History: From Then to Now”. Shelf. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://shelf.io/blog/master-nlp-history-from-then-to-now/
10. purpleSlate. “Evolution of NLP: From Past Limitations to Modern Capabilities”. Medium. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://medium.com/@social_65128/evolution-of-nlp-from-past-limitations-to-modern-capabilities-6dc1505faeb6

11. IJISET | International Journal of Innovative Science, Engineering and Technology. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.ijiset.com/v1s3/IJISET_V1_I3_92.pdf
12. M. G. Muthee, Mutua Makau, and Omamo Amos, "A review of techniques for morphological analysis in natural language processing", AJSTSS, vol. 1, no. 2, pp. 93–103, Dec. 2022. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://journals.must.ac.ke/index.php/AJSTSS/article/view/11/114
13. "Data Science: Natural Language Processing (NLP)". Oak-Tree Technologies. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.oak-tree.tech/blog/data-science-nlp
14. "Syntactic Analysis: A Power Tool In NLP Made Easy With Examples, Illustrations & Tutorials". Spot Intelligence. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://spotintelligence.com/2023/10/28/syntactic-analysis-nlp/
15. "Semantic Analysis: What Is It, How & Where To Works". QuestionPro. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.questionpro.com/blog/semantic-analysis/
16. T. Mathur. "Pragmatics in NLP - Scaler Topics". Scaler Topics. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.scaler.com/topics/nlp/pragmatics-in-nlp/
17. Учасники проектів Вікімедіа. "Вінцюк Тарас Климович — Вікіпедія". Вікіпедія. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://uk.wikipedia.org/wiki/Вінцюк_Тарас_Климович
18. "Історія ІПШІ". Інститут проблем штучного інтелекту (Київ, Україна). Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.ipai.net.ua/uk/istoriya-ipshi
19. "UNLP 2025 | History". UNLP 2025 | The Fourth Ukrainian Natural Language Processing Workshop. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://unlp.org.ua/history/
20. Галактика Слова. Галині Макарівні Гнатюк / Ін-т укр. мови НАН України. – К. : Вид. дім Дмитра Бураго, 2020. – С. 135–141. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.researchgate.net/publication/344842033_Velikij_elektronnij_slovnik_ukrainskoi_movi_VESUM_ak_zasib_NLP_dla_ukrainskoi_movi_Galaktika_Slova_Galini_Makarivni_Gnatuk/link/5fa110cd458515b7cfb5cc97/download?_tp=eyJjb250ZXh0Ijp7InBhZ2UiOiJwdWJsaWNhdGlvbiIsInByZXZpb3VzUGFnZSI6bnVsbH19

21. Комп’ютерна лінгвістика: сучасне та майбутнє. Матеріали Міжнародної науково-практичної конференції – К.: КНЛУ, 2012.– 52 с. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: http://www.mova.info/zbirnyk.pdf
22. “GitHub - brown-uk/corpus: Браунський корпус української мови”. GitHub. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://github.com/brown-uk/corpus
23. “БрУК”. Facebook. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://www.facebook.com/BrUKgroup
24. “ГРАК - site.name”. ГРАК - site.name. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://uacorpus.org
25. “GitHub - brown-uk/nlp_uk: This is a project to demonstrate NLP API from LanguageTool for Ukrainian language.” GitHub. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://github.com/brown-uk/nlp_uk
26. “GitHub - osyvokon/awesome-ukrainian-nlp: Curated list of Ukrainian natural language processing (NLP) resources (corpora, pretrained models, libriaries, etc.)”. GitHub. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://github.com/osyvokon/awesome-ukrainian-nlp
27. Анатолій Загнітко, Ілля Данилюк, Жанна Краснобаєва-Чорна, Оксана Путіліна, Ганна Ситар. Парадигмально-категорійні основи прикладної лінгвістики : Монографія. – Вінниця : «ТОВ Нілан-ЛТД», 2015. – 472 с.ISBN 978-617-7212-94-1. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://ulif.mon.gov.ua/system/files/prikladna_mono_2015.pdf
28. Надя Осмокеску. “Що заважає розвитку NLP”. robot_dreams - онлайн-курси для фахівців у сфері big data, machine learning, data science | Робот Дрімс. Дата звернення: 8 трав. 2025. [Онлайн]. Доступно: https://robotdreams.cc/uk/blog/87-chto-meshaet-razvitiyu-nlp