УДК 004.934:[81:004.656(477)
DOI: 10.31866/2617-796X.8.2.2025.347880

Костянтин Ткаченко,
кандидат економічних наук, доцент,
доцент кафедри інформаційних технологій,
Навчально-науковий інститут управління,
технологій та правових наук,
Національний транспортний університет,
Київ, Україна,
доцент кафедри інженерії програмного забезпечення,
Державний університет «Київський авіаційний інститут»,
Київ, Україна
tkachenko.kostyantyn@gmail.com
https://orcid.org/0000-0003-05493396

Олександр Смирнов,
магістрант кафедри інформаційних технологій,
Навчально-науковий інститут управління,
технологій та правових наук,
Національний транспортний університет,
Київ, Україна
juvenilereader@gmail.com
https://orcid.org/0009-0005-4530-9700

МОДЕЛЮВАННЯ ПРОЦЕСІВ ЛІНГВІСТИЧНОГО АНАЛІЗУ УКРАЇНСЬКОМОВНИХ ТЕКСТІВ

На сьогодні широко застосовують автоматизацію обробки текстів природною мовою в різних сферах (державному секторі, науці, освіті, медіа, повсякденних сервісах тощо). Це потребує відповідних програмних інструментів (сервісів, засобів), які будуть спроможні забезпечувати таку автоматизовану обробку.

Метою статті є аналіз і дослідження проблем моделювання процесів лінгвістичного аналізу українських природномовних текстів і проєктування відповідного програмного забезпечення з визначенням функціональних можливостей його компонентів.

Методами дослідження є методи порівняльного аналізу основних програмних рішень цієї предметної області (лінгвістичний аналіз текстів, що представлені природною мовою), систематизації підходів до процесів автоматизованої обробки текстів та формалізації цих процесів у вигляді відповідної процесної моделі.

Новизною проведеного дослідження є аналіз сучасних проблем систем підтримки процесів автоматизованого оброблення текстів природною мовою, зокрема українською; розробка процесної моделі, що поєднує різні етапи та фази лінгвістичного аналізу текстів і проєктування відповідного програмного рішення підтримки цієї моделі.

© Ткаченко К. О.,
© Смирнов О. О.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Висновки. У роботі досліджено основні проблеми обробки природномовних текстів; визначено основні методи обробки українськомовних текстів; проведено аналіз і систематизацію сучасних систем, що підтримують окремі етапи автоматизованої обробки природномовних текстів; описано проєкт авторського програмного рішення що забезпечуватиме лінгвістичний аналіз українськомовних текстів; запропоновано процесну модель лінгвістичного аналізу природномовних текстів. Використання запропонованої процесної моделі з боку користувачів та розробників сприятиме полегшенню розгортання якісних українськомовних сервісів; з боку установ (відповідних програмних продуктів) під час модифікації системи лінгвістичного аналізу сприятиме отриманню більш об'єктивного уявлення про найчастіші проблеми лінгвістичного аналізу (так звані мовні проблеми), прогалини лексикону та пріоритети оновлень українськомовного NLP.

Ключові слова: природномовні тексти; лінгвістичний аналіз; моделювання процесів лінгвістичного аналізу; процесна модель; морфологічний словник VESUM; Universal Dependencies; розомонімізація; токенізація.

Постановка проблеми. Стрімке зростання обсягів українськомовних текстів у державних сервісах, освіті, медіа та бізнесі обумовило розробку програмних рішень (систем, сервісів, застосунків), призначених для автоматизації опрацювання текстів природною мовою. Завдання таких програмних рішень полягають в якісному та надійному аналізі великих масивів українськомовних текстів у обраних предметних галузях професійного спілкування.

Реалізація вказаних завдань потребує від систем оброблення природної мови не лише високої точності, а й відтворюваності, сумісності та майже «промислової» надійності.

Наявні ж програмні рішення здебільшого:

- зосереджені на окремих етапах лінгвістичного аналізу (наприклад, таких як токенізація (lang-uk/tokenize-uk, n.d.), лематизація (Старко та Рисін, 2020; Starko and Rysin, 2022), PoS-тегування (UDPipe 2 Models, n.d.), парсинг (Available Models & Languages, n.d.);
- фрагментарні, які по-різному трактують граматичні ознаки та не забезпечують узгодженого «наскрізного» процесу.

Усе вищевказане призводить до:

- накопичення помилок, які виникають у процесі переходу між етапами (фазами) лінгвістичного аналізу;
- ускладнення повторного використання результатів аналізу;
- зниження якості пошуку (лінгвістичних одиниць, слів, лексем тощо);
- ускладнення щодо:
  - надання граматичних підказок;
  - розпізнавання іменованих сутностей (*Named Entity Recognition*, *NER*) (Available Models & Languages, n.d.);
  - синтаксичного аналізу в реальних застосунках.

На практиці і користувачі, і розробники частіше стикаються з розрізненими інструментами лінгвістичного аналізу текстів чи інструментами, що початково були

розроблені під моделі лінгвістичного аналізу англомовних текстів, а потім адаптовані під україномовні тексти.

Усе це згодом виявляє невідповідність наявних програмних рішень та відповідних моделей лінгвістичного аналізу потребам користувачів і розробників. Ці невідповідності обумовлюють, зокрема:

- похибки лематизації та PoS-тегування;
- неузгодженість грамем;
- втрату якості синтаксичного аналізу;
- нерівномірне покриття так званих дефісованих форм, апострофа, варіантів правопису й новотворів.

Отже, актуальність розробки цілісної процесної моделі лінгвістичного аналізу українських текстів не викликає сумнівів.

Аналіз останніх досліджень і публікацій. За останні 5–7 років у вітчизняному досвіді обробки природної мови (*Natural Language Processing*, *NLP*) (Старко та Рисін, 2020; Starko and Rysin, 2022; Haltiuk and Smywiński-Pohl, 2024; Starko and Rysin, 2023) сформувалася ресурсно-процесна база для моделювання повного конвеєра лінгвістичного аналізу українських текстів. Наприклад, VESUM (Старко та Рисін, 2020; Starko and Rysin, 2022; Starko, Rysin and Shvedova, 2021) – морфологічний словник з так званим динамічним тегуванням, що підтримує аналіз, зокрема:

- гібридних / продуктивних утворень;
- «дефісованих» форм;
- прислівників.

VESUM інтегрований у виробничі сервіси (Apache Lucene, орфографічні та граматичні рушії) і доступний під відкритими ліцензіями даних та інструментів. Усе це закриває базові етапи конвеєра лінгвістичного аналізу текстів (такі як лематизація, PoS, грамеми) та задає формат API для подальших модулів обробки (Старко та Рисін, 2020; Starko and Rysin, 2022).

Стандартизацію виходів процесу забезпечує екосистема Universal Dependencies (UD Ukrainian IU, n.d.; Universal Dependencies/UD_Ukrainian-IU, n.d.; UD Ukrainian ParlaMint, n.d.), яка містить:

- базовий UD_Ukrainian-IU (правові, новинні, художні, Вікі-домени);
- UD_Ukrainian-ParlaMint (парламентський корпус).

Ці ресурси фіксують уніфікований простір ознак (UPOS/FEATS/DEPREL) і дають змогу сумісно тренувати й оцінювати тегери та парсери, що є критичним для узгодженого pipeline «морфологія → синтаксис» (UDPipe 2 Models, n.d.; Available Models & Languages, n.d.). Цей pipeline передбачає послідовність кроків щодо створення програмного продукту для лінгвістичного аналізу текстів природною мовою (pipeline візуалізується, щоб стандартизувати, керувати й оптимізувати робочий процес).

На інструментальному рівні підтримки автоматизації лінгвістичного аналізу текстів доступні готові ланцюжки:

- токенізатор tokenize-uk (обробка апострофа + обробка дефісів);
- UDPipe 2 (токенізація + PoS + лематизація + депенденсі-парсинг у одному виклику, використовуються моделі для української мови) (UDPipe 2 Models, n.d.);

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

– Stanza (UD-сумісні моделі).

Усі вказані ланцюжки природно «вкладаються» в процесну модель (lang-uk/tokenize-uk, n.d.; UDPipe 2 Models, n.d.; Available Models & Languages, n.d.; Starko and Rysin, 2023): нормалізація → токенізація → морфологія → розомонімізація → синтаксис.

Виробничі інтеграції ресурсів і компонентів VESUM/UD підтверджують зрілість підходу:

– український модуль LanguageTool використовує Morfologik-спелер (Morfologik speller for Ukrainian, 2018), побудований на основі VESUM;

– для VESUM створено окремий пошуковий рушій (eLex 2023) (Krashtan, 2023);

– морфологічний аналізатор на базі VESUM застосовано для переіндексації української Вікіпедії.

Усе це свідчить про практичну сумісність між словником, пошуком і редакторськими сервісами (Starko and Rysin, 2022; Morfologik speller for Ukrainian, 2018; Krashtan, 2023). Паралельно з класичним модульним конвеєром VESUM/UD розвивається мономовне моделювання:

– LiBERTa Large (BERT-Large, навчена з нуля на українських текстах) підвищує якість NLU-завдань (Haltiuk and Smywiński-Pohl, 2024);

– численні роботи демонструють ефективність донавчання BERT/DistilBERT/XLM-R на українських корпусах (Prytula, 2024).

У межах цієї процесної моделі йдеться про два рівнозначні архітектурні шляхи розвитку NLP-конвеєра для української мови – з різними компромісами, але зі спільним інтерфейсом, тобто дві траєкторії його архітектурної еволюції:

– класичний модульний конвеєр з VESUM/UD;

– гібрид, коли розомонімізація (Starko and Rysin, 2023) / семантика делегуються трансформерам, а входи стандартизуються морфологічним шаром (Haltiuk and Smywiński-Pohl, 2024; Prytula, 2024).

Динамічна розомонімізація висвітлюється деякими авторами у зв’язці з VESUM і корпусами (GRAC, ParlaMint) (Starko and Rysin, 2023; Starko, Rysin and Shvedova, 2021; ANN: LanguageTool 6.4, n.d.):

– тегери з компонентом dynamic tagging досягають високого покриття OOV (Starko and Rysin, 2022);

– регулярне добудовування реєстру знижує розрив між словником і реальними текстами (Старко та Рисін, 2020).

Розглянутий напрям досліджень підсилює етап ланцюга «морфологія → контекст» (Starko and Rysin, 2022; UD Ukrainian ParlaMint, n.d.; Starko and Rysin, 2023).

Мета статті. Метою є аналіз і дослідження проблем моделювання процесів лінгвістичного аналізу українських природномовних текстів та проєктування відповідного програмного забезпечення з визначенням функціональних можливостей його компонентів.

Досягнення мети передбачає виконання таких завдань:

– визначення основних проблем лінгвістичного аналізу природномовних текстів;

– визначення основних методів лінгвістичного аналізу українськомовних текстів (зокрема, професійного спрямування);

– проведення аналізу та систематизації наявних програмних рішень, що підтримують проведення окремих етапів лінгвістичного аналізу природномовних текстів;

– визначення вимог до програмного рішення, що буде забезпечувати лінгвістичний аналіз українськомовних текстів (зокрема, професійного спрямування);

– проведення проєктування авторського програмного рішення що буде забезпечувати лінгвістичний аналіз українськомовних текстів.

Виклад основного матеріалу дослідження. Специфіка української мови (багата словозміна, апостроф та дефісовані форми, варіативність правопису, вільний порядок слів, активні новотвори й іншомовні вкраплення) загострює вимоги до формального опису й узгодження етапів оброблення.

Відсутність уніфікованої процесної моделі (із чітко визначеними входами / виходами, інтерфейсами між модулями, правилами розомонімізації та стандартизованим відображенням ознак (на кшталт UD)) робить результати різних систем лінгвістичного аналізу несумісними, унеможливлює прозоре тестування якості та гальмує впровадження у виробництві (*batch* і *streaming*) (UD Ukrainian IU, n.d.; Universal Dependencies/UD_Ukrainian-IU, n.d.; UD Ukrainian ParlaMint, n.d.; lang-uk/tokenize-uk, n.d.; UDPipe 2 Models, n.d.; Available Models & Languages, n.d.).

Отже, науково-практичне завдання цієї роботи полягає в моделюванні процесів лінгвістичного аналізу українських природномовних текстів як цілісного «конвеєра»:

– формалізувати послідовність і взаємодію етапів: нормалізація → токенізація → морфологічний аналіз/синтез → контекстна розомонімізація → синтаксично-семантичні представлення;

– визначити уніфіковані формати даних та API;

– забезпечити відтворювані метрики на кожному кроці;

– підтримати динамічне покриття новотворів і доменної лексики, а також сумісність зі стандартами корпусної розмітки.

Результатом має стати відтворювана, масштабована і взаємосумісна (тобто модулі взаємозамінні й коректно працюють разом завдяки уніфікованим форматам I/O – JSON, CoNLL-U, стандартним API та відповідності UD) процесна модель, що зменшує похибки на базових етапах і підвищує якість кінцевих сервісів. Без процесної моделі складно одночасно досягти відтворюваності, сумісності та виробничої надійності. Тому варто базуватися на зв'язці машиночитного лексикону VESUM і корпусних стандартів UD, які задають словотвірно-морфологічну та синтаксичну «мову» взаємодії модулів конвеєра (Старко та Рисін, 2020; Starko and Rysin, 2022; UD Ukrainian IU, n.d.; Universal Dependencies/UD_Ukrainian-IU, n.d.; UD Ukrainian ParlaMint, n.d.).

Розроблення програмного забезпечення та формалізація процесної моделі лінгвістичного аналізу українськомовних природномовних текстів забезпечує узгоджену послідовність етапів: нормалізація → токенізація → морфологічний аналіз / синтез (з опорою на лексикографічні ресурси) → контекстна розомонімізація → синтаксично-семантичні представлення (UD-сумісні).

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Модель має враховувати специфіку української мови (багата словозміна, вільний порядок слів, дефіс / апостроф, іншомовні вкраплення), підтримувати стандартизовані формати I/O та API, забезпечувати відтворювані метрики на кожному кроці та масштабування для пакетної й потокової обробки.

На вході pipeline авторський програмний продукт отримує так званий «сирий» текст, який спершу нормалізується (здійснюється уніфікація Unicode, знаходяться: символ «’» (апостроф), дефіси, літера «ґ», символ « «» » (лапки)), а далі текст токенізується, тобто відбувається поділ на речення та індивідуальні значення).

Для української мови етап нормалізації й токенізації – це джерело системних похибок. Помилки на цьому рівні конвеєра мультиплікуються на всіх наступних кроках. На практиці більш доцільно спиратися на tokenize-uk (відкритий токенізатор; ліцензія MIT), що акуратно працює з апострофом чи дефісом і реченнєвими межами, знижуючи шум на вході відповідної морфології (lang-uk/tokenize-uk, n.d.; Dyomkin and Chaplinsky, 2017). Деталізацію цього конвеєра показано на рис. 1.

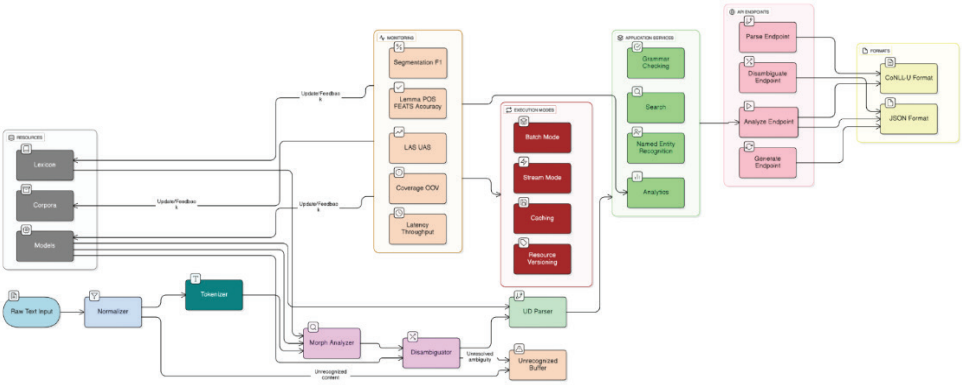

Рис. 1. Узагальнена архітектура процесної моделі
Джерело: авторська розробка

Конвеєр починається з Raw Text Input, який проходить через Normalizer (уніфікація Unicode, апостроф / дефіси, ґ, лапки) та Tokenizer (поділ на речення і токени з офсетами). Далі Morph Analyzer на базі VESUM виконує словниковий пошук, застосовує парадигми, «динамічне тегування» продуктивних форм і OOV-гіпотези. Disambiguator поєднує правила узгодження з легкою ML-моделлю та повертає 1-best {lemma, POS, FEATS}.

UD Parser відображає ознаки у формат UD (UPOS/XPOS/FEATS/DEPREL) і формує вихід у JSON/CoNLL-U, який споживають прикладні сервіси: Grammar Checking, Search (лем- та словоформна індексація), NER і Analytics. Невпізнані чи конфліктні випадки накопичує Unrecognized Buffer. З нього оновлення повертаються у блок Resources: Lexicon (VESUM), Corpora (UD_Ukrainian-IU, ParlaMint тощо) та Models. Так замикається цикл якості. Виконанням керує блок Execution Modes (batch/

stream, кеші, версіонування ресурсів), а стан системи відстежує Monitoring за ключовими метриками: Segmentation F1, Lemma/POS/FEATS accuracy, Coverage OOV, LAS/UAS, Latency/Throughput (p95/p99). Доступ назовні забезпечують версійовані API-ендпоїнти: /analyze, /disambiguate, /parse, /generate. Усе це утворює єдиний, відтворюваний і масштабований процес від «сирого» тексту до прикладних сервісів із замкненим зворотним зв'язком оновлень. Центром (ядром, так званим серцем) моделі є морфологічний шар на VESUM. Лексикон надає леми, класи словозміни, правила афіксації та динамічне тегування продуктивних моделей (типу *по-Х-ськи*, «дефісовані» форми, новотвори), забезпечуючи високе покриття реальних українськомовних текстів і слугуючи основою (так званим хребтом) для подальших етапів лінгвістичного аналізу (Старко та Рисін, 2020; Starko and Rysin, 2022; Haltiuk and Smywiński-Pohl, 2024).

У виробничих (тобто реальних прикладних / «продакшн»-сервісах, а не лабораторних прототипах) сценаріях машиночитний лексикон VESUM разом із морфологічним аналізатором / генератором уже застосовано як спелер і лематизатор у LanguageTool та для лем-індексації – в Lucene/Elastic. Деякі автори (Starko and Rysin, 2022; Morfologik speller for Ukrainian, 2018; Krashtan, 2023) розглядають спеціалізовані пошукові інтерфейси до словника, важливі для контролю якості та еволюції реєстру.

Хід однієї словоформи через аналізатор та синтезатор VESUM проілюстровано на рис. 2. Потік починається із «сирого» тексту, який нормалізується (Unicode, апостроф, дефіси, ґ, лапки) і токенізується на речення та токени. Далі токен проходить морфологічний аналіз на базі VESUM: словниковий lookup і FST-правила генерують n-best інтерпретації, за потреби підключається «динамічне тегування» та OOV-гіпотези, після чого спрацьовує фільтр узгодженості ознак. Блок розомонімізації поєднує правила узгодження з легкою ML-моделлю й обирає 1-best {lemma, POS, FEATS}. Результат мапується у стандарт UD (UPOS/XPOS/FEATS/DEPREL) і віддається як JSON/CoNLL-U. Паралельно модуль синтезу з леми та грамем генерує поверхневу форму та виконує round-trip-перевірку (analyze∘generate), що валідує (перевіряє) узгодженість аналізу. Отримані подання споживають прикладні сервіси (пошук, перевірка, NER); невизначені випадки фіксуються для подальшого донавчання й оновлення ресурсів.

Невід'ємним кроком процесної моделі (конвеєра лінгвістичного аналізу) є контекстна розомонімізація (Starko and Rysin, 2023): поєднання правил узгодження (рід, число, відмінок) із легкою ML-моделлю (CRF/BiLSTM-CRF або компактний трансформер), яка споживає ознаки VESUM. Досвід створення POS-«золота» (ідеться про золотий стандарт – вручну / експертно розмічений еталонний корпус частин мови) для української мови показує, що така зв'язка підвищує і покриття, і якість виявлення помилок, а критичним фактором стабільної дизамбіґуації в потоці (Starko and Rysin, 2022; Haltiuk and Smywiński-Pohl, 2024) є саме гібридний підхід «правила узгодження + легка ML-модель» на ознаках VESUM.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

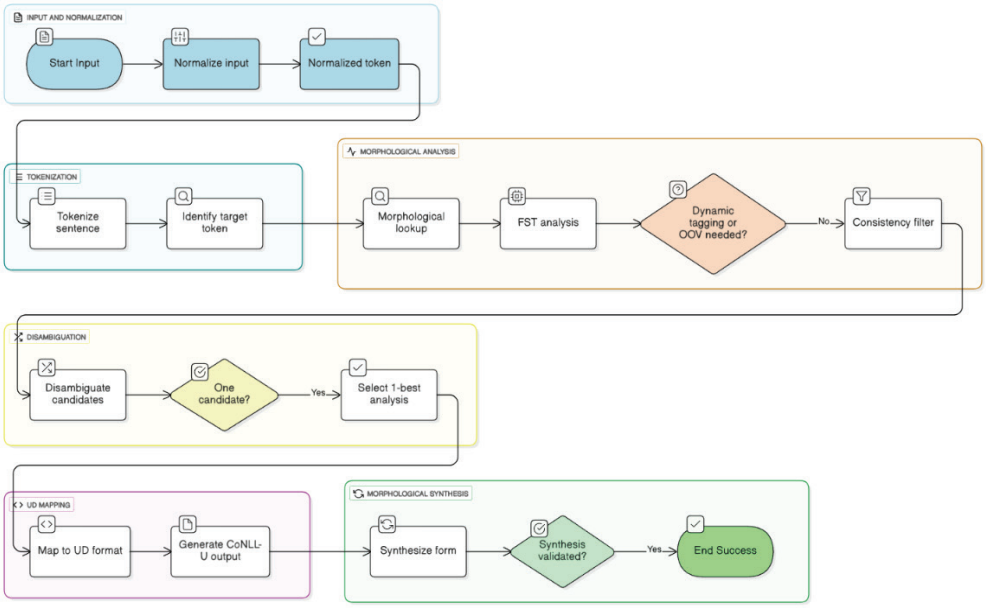

Рис. 2. Pipeline через VESUM-аналізатор і синтезатор
Джерело: авторська розробка

Результати морфологічного аналізу далі відображаються у стандарті Universal Dependencies (UPOS/FEATS/XPOS/DEPREL) (UD Ukrainian IU, n.d.; Universal Dependencies/UD_Ukrainian-IU, n.d.; UD Ukrainian ParlaMint, n.d.), що дає змогу:

– використовувати готові парсери (UDPipe 2, Stanza) (UDPipe 2 Models, n.d.; Available Models & Languages, n.d.);

– переносити моделі між доменами (предметними областями, тексти з яких аналізуються) без так званих «клейових» конвертерів (UD Ukrainian IU, n.d.; Universal Dependencies/UD_Ukrainian-IU, n.d.; UD Ukrainian ParlaMint, n.d.; lang-uk/tokenize-uk, n.d.; UDPipe 2 Models, n.d.; Available Models & Languages, n.d.).

Практичний сенс такого відображення – не тільки уніфікація форматів, а й зменшення технічного боргу, тобто накопичених архітектурних компромісів і ручних засобів розробки, що ускладнюють інтеграцію, тестування та розвиток системи. У підсумку дані (корпуси, словники, анотації), моделі (навчені алгоритми / ваги для задач), і бенчмарки (стандартизовані тести та метрики порівняння) починають «говорити однією мовою».

Застосування процесної моделі в реальних програмних сервісах слугує перевіркою її життєздатності. Інформаційно-пошукові системи для документів виграють від поєднання лем-індексації та індексації словоформ, що підвищує стійкість до словозміни (практичні кейси охоплюють переіндексацію української Вікіпедії).

Перевірка та редагування тексту покладаються на спелер і граматичні правила, які спираються на лематизацію й тегування з VESUM; спільнота LanguageTool публікує відкриті обговорення продуктивності та регулярні оновлення правил для української мови. Лексикографічний пошук і вебверсії словника замикають цикл якості (виявлення прогалин у пошуку / аналізі) лексикону VESUM та мовного конвеєра: користувацький і корпусний зворотний зв'язок (помилки, OOV, частотні списки) оперативно повертається в реєстр, після чого оновлюються правила / моделі й перевидається реліз (Universal Dependencies/UD_Ukrainian-IU, n.d.; Krashtan, 2023).

У підсумку маємо не «набір утиліт», а процес із замкненим зворотним зв'язком, коли ресурси (лексикон, моделі, мовні корпуси) версіонуються й еволюціонують разом із пайплайном. Оцінювання NLP-конвеєра ведеться не списком ізольованих метрик, а як моніторинг процесу наскрізної обробки (послідовний pipeline: від «сирого» тексту):

- на вході – *segmentation F1*;
- під час морфологічного аналізу – *lemma/POS/FEATS accuracy* та *coverage/OOV*;
- під час синтаксичного аналізу – *LAS/UAS* за бенчмарками UD.

Додатково відстежуються латентність (p95/p99) і пропускна здатність у batch- / stream-режимах; орієнтиром слугують практики відкритих UD-пайплайнів на кшталт UDPipe 2 та Stanza, а також протоколи оцінювання за українським POS «золотим стандартом» (UNLP 2023). Саме так метрики виконують свою головну функцію – керують рішеннями про конфігурацію конвеєра («правила ↔ ML ↔ гібрид») під конкретний домен, а не просто підсумовують відсотки.

Використання розробленої авторської процесної моделі моделювання лінгвістичного аналізу українськомовних текстів дає змогу, зокрема:

- забезпечити для реальних текстових потоків узгоджений і відтворюваний конвеєр: нормалізація → токенізація → морфологія → розомонімізація → UD;
- створити простий зв'язок між користувачем і системою через анотації та стандартизовані API (JSON, CoNLL-U);
- не вимагати від прикладних команд глибоких лінгвістичних знань завдяки зрозумілим профілям застосування й завдяки автоналаштуванню піддоменів;
- генерувати кілька конфігурацій процесу під різні домени (новини, держдокументи, розмовні дані) з контрольованими метриками якості;
- підвищити якість пошуку, перевірки тексту, NER і парсингу за рахунок VESUM, «динамічного тегування» та сумісності з UD;
- вбудувати прозорий моніторинг (SegF1, Lemma/POS/FEATS, LAS/UAS, OOV, latency/throughput) і цикл зворотного зв'язку для оновлення лексикону / правил / моделей.

Використання розробленої авторської процесної моделі та відповідного програмного забезпечення дасть змогу, зокрема:

- надавати зручні інструменти для конфігурації та валідації конвеєра;
- установити прозорий зв'язок між користувачем (аналітиком, розробником, лінгвістом) і системою;
- не вимагати глибоких лінгвістичних знань від прикладних команд завдяки зрозумілим параметрам та профілям застосування;

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

– генерувати декілька варіантів / конфігурацій процесу піддоменів (наприклад, таких як новини, державні документи, розмовні дані тощо);

– підвищити якість пошуку, граматичних підказок, NER і парсингу завдяки уніфікованим ознакам і сумісності з корпусними стандартами;

– спростити інтеграцію в наявні IT-ландшафти через стандартизовані API та метрики.

Висновки. У роботі досліджено та визначено основні проблеми обробки природномовних текстів, крім того:

– визначено основні методи обробки українськомовних текстів;

– проведено аналіз і систематизацію сучасних систем, що підтримують окремі етапи автоматизованої обробки природномовних текстів;

– описано проєкт авторського програмного рішення що забезпечуватиме лінгвістичний аналіз українськомовних текстів;

– запропоновано процесну модель лінгвістичного аналізу природномовних текстів.

Використання запропонованої процесної моделі:

– з боку користувачів та розробників сприятиме полегшенню розгортання якісних українськомовних сервісів;

– з боку установ (відповідних програмних продуктів) під час модифікації системи лінгвістичного аналізу (зі зворотним зв'язком від реальних текстових потоків) сприятиме отриманню більш об'єктивного уявлення про найчастіші проблеми лінгвістичного аналізу (так звані мовні проблеми), проблеми, прогалини лексикону та пріоритети оновлень українськомовного NLP.

СПИСОК ПОСИЛАНЬ

Старко, В.Ф. та Рисін, А., 2020. Великий електронний словник української мови (ВЕСУМ) як засіб NLP для української мови. В: *Галактика Слова. Галині Макарівні Гнатюк*. Київ: Видавничий дім Дмитра Бураго, с.135-141. [online] Доступно: <https://www.researchgate.net/publication/344842033_Velikij_elektronnij_slovnik_ukrainskoi_movi_VESUM_ak_zasib_NLP_dla_ukrainskoi_movi_Galaktika_Slova_Galini_Makarivni_Gnatuk> [Дата звернення 24 вересня 2025].

ANN: LanguageTool 6.4, n.d. *LT*. [online] Available at: <https://forum.languagetool.org/t/ann-languagetool-6-4/9950> [Accessed 25 September 2025].

Available Models & Languages, n.d. *Stanza*, [online] Available at: <https://stanfordnlp.github.io/stanza/available_models.html> [Accessed 24 September 2025].

Dyomkin, V. and Chaplinsky, D., 2017. *Tokenize UK Documentation. Release 0.2.0.* [online] Available at: <https://tokenize-uk.readthedocs.io/_/downloads/en/latest/pdf/> [Accessed 24 September 2025].

Haltiuk, M. and Smywiński-Pohl, A., 2024. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. In: *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024"*, Torino, Italia, May 25, 2024. [online]

Torino: ELRA Language Resources Association, pp.120-128. Available at: <https://aclanthology.org/2024.unlp-1.14.pdf> [Accessed 25 September 2025].
Krashtan, T., 2023. A Search Engine for the Large Electronic Dictionary of the Ukrainian Language (VESUM). In: *Electronic Lexicography in the 21st Century*. The eighth eLex conference, Brno, Czech Republic, June 27-29, 2023. [online] Brno, pp.308-321. Available at: <https://elex.link/ojs/index.php/elex/article/view/33> [Accessed 25 September 2025].
lang-uk/tokenize-uk, n.d. *GitHub*. [online] Available at: <https://github.com/lang-uk/tokenize-uk> [Accessed 30 September 2025].
Morfologik speller for Ukrainian, 2018. *LT*, [online] July. Available at: <https://forum.languagetool.org/t/morfologik-speller-for-ukrainian/3188> [Accessed 25 September 2025].
Prytula, M., 2024. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence,* [e-journal] 2, pp.85-97. https://doi.org/10.15407/jai2024.02.085
Starko, V. and Rysin, A., 2022. VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*. Vol. I: Main Gliwice, Poland, May 12-13, 2022. [online] CEUR-WS.org, online, pp.61-70. Available at: <https://ceur-ws.org/Vol-3171/paper8.pdf> [Accessed 24 September 2025].
Starko, V. and Rysin, A., 2023. Creating a POS Gold Standard Corpus of Modern Ukrainian. In: *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, Dubrovnik, Croatia, May 5, 2023. [online] Dubrovnik: Association for Computational Linguistics, pp.91-95. Available at: <https://aclanthology.org/2023.unlp-1.11.pdf> [Accessed 25 September 2025].
Starko, V., Rysin, A. and Shvedova, M., 2021. Ukrainian Text Preprocessing in GRAC. In: *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, September 22-25, 2021. [online] Institute of Electrical and Electronics Engineers, Vol.2, pp.101-104. Available at: <https://www.researchgate.net/publication/357362512_Ukrainian_Text_Preprocessing_in_GRAC> [Accessed 25 September 2025].
UD Ukrainian IU, n.d. [online] Available at: <https://universaldependencies.org/treebanks/uk_iu/> [Accessed 24 September 2025].
UD Ukrainian ParlaMint, n.d. *UD_Ukrainian-ParlaMint.* [online] Available at: <https://universaldependencies.org/treebanks/uk_parlamint/> [Accessed 24 September 2025].
UDPipe 2 Models, n.d. *Institute of Formal and Applied Linguistics.* [online] Available at: <https://ufal.mff.cuni.cz/udpipe/2/models> [Accessed 30 September 2025.
Universal Dependencies/UD_Ukrainian-IU, n.d. *GitHub*. [online] Available at: <https://github.com/UniversalDependencies/UD_Ukrainian-IU> [Accessed 24 September 2025].

REFERENCES

ANN: LanguageTool 6.4, n.d. *LT*. [online] Available at: <https://forum.languagetool.org/t/ann-languagetool-6-4/9950> [Accessed 25 September 2025].
Available Models & Languages, n.d. *Stanza*, [online] Available at: <https://stanfordnlp.github.io/stanza/available_models.html> [Accessed 24 September 2025].
Dyomkin, V. and Chaplinsky, D., 2017. Tokenize UK Documentation. *Release 0.2.0.* [online] Available at: <https://tokenize-uk.readthedocs.io/_/downloads/en/latest/pdf/> [Accessed 24 September 2025].

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

Haltiuk, M. and Smywiński-Pohl, A., 2024. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. In: *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024»*, Torino, Italia, May 25, 2024. [online] Torino: ELRA Language Resources Association, pp.120-128. Available at: <https://aclanthology.org/2024.unlp-1.14.pdf> [Accessed 25 September 2025].

Krashtan, T., 2023. A Search Engine for the Large Electronic Dictionary of the Ukrainian Language (VESUM). In: *Electronic Lexicography in the 21st Century. The eighth eLex conference*, Brno, Czech Republic, June 27-29, 2023. [online] Brno, pp.308-321. Available at: <https://elex.link/ojs/index.php/elex/article/view/33> [Accessed 25 September 2025].

lang-uk/tokenize-uk, n.d. *GitHub*. [online] Available at: <https://github.com/lang-uk/tokenize-uk> [Accessed 30 September 2025].

Morfologik speller for Ukrainian, 2018. *LT*, [online] July. Available at: <https://forum.languagetool.org/t/morfologik-speller-for-ukrainian/3188> [Accessed 25 September 2025].

Prytula, M., 2024. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence*, [e-journal] 2, pp.85-97. https://doi.org/10.15407/jai2024.02.085

Starko, V. and Rysin, A., 2022. VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*. Vol. I: Main Gliwice, Poland, May 12-13, 2022. [online] CEUR-WS.org, online, pp.61-70. Available at: <https://ceur-ws.org/Vol-3171/paper8.pdf> [Accessed 24 September 2025].

Starko, V. and Rysin, A., 2023. Creating a POS Gold Standard Corpus of Modern Ukrainian. In: *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, Dubrovnik, Croatia, May 5, 2023. [online] Dubrovnik: Association for Computational Linguistics, pp.91-95. Available at: <https://aclanthology.org/2023.unlp-1.11.pdf> [Accessed 25 September 2025].

Starko, V., Rysin, A. and Shvedova, M., 2021. Ukrainian Text Preprocessing in GRAC. In: *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, September 22-25, 2021. [online] Institute of Electrical and Electronics Engineers, Vol.2, pp.101-104. Available at: <https://www.researchgate.net/publication/357362512_Ukrainian_Text_Preprocessing_in_GRAC> [Accessed 25 September 2025].

Starko, V.F. and Rysin, A., 2020. Velykyi elektronnyi slovnyk ukrainskoi movy (VESUM) yak zasib NLP dlia ukrainskoi movy [The Great Electronic Dictionary of the Ukrainian Language (VESUM) as a NLP tool for the Ukrainian language]. In: *Halaktyka Slova. Halyni Makarivni Hnatiuk* [Galaktika Slova. Galina Makarivna Gnatyuk]. Kyiv: Vydavnychyi dim Dmytra Buraho, pp.135-141. [online] Dostupno: <https://www.researchgate.net/publication/344842033_Velikij_elektronnij_slovnik_ukrainskoi_movi_VESUM_ak_zasib_NLP_dla_ukrainskoi_movi_Galaktika_Slova_Galini_Makarivni_Gnatuk> [Data zvernennia 24 veresnia 2025].

UD Ukrainian IU, n.d. [online] Available at: <https://universaldependencies.org/treebanks/uk_iu/> [Accessed 24 September 2025].

UD Ukrainian ParlaMint, n.d. *UD_Ukrainian-ParlaMint*. [online] Available at: <https://universaldependencies.org/treebanks/uk_parlamint/> [Accessed 24 September 2025].

UDPipe 2 Models, n.d. *Institute of Formal and Applied Linguistics*. [online] Available at: <https://ufal.mff.cuni.cz/udpipe/2/models> [Accessed 30 September 2025.

Universal Dependencies/UD_Ukrainian-IU, n.d. *GitHub*. [online] Available at: <https://github.com/UniversalDependencies/UD_Ukrainian-IU> [Accessed 24 September 2025].

UDK 004.934:[81:004.656(477)

Kostiantyn Tkachenko,
PhD in Economics, Associate Professor,
Associate Professor at the Department of Information Technologies
Educational and Scientific Institute of Management,
Technology and Legal Sciences,
National Transport University,
Kyiv, Ukraine
Associate Professor at the Department of Software Engineering
State University "Kyiv Aviation Institute",
Kyiv, Ukraine
tkachenko.kostyantyn@gmail.com
ORCID ID: 0000-0003-0549-3396

Smyrnov Oleksandr,
Master's Student at the Department of Information Technologies,
Educational and Scientific Institute of Management, Technology and Legal Sciences,
National Transport University,
Kyiv, Ukraine
juvenilereader@gmail.com
ORCID ID: 0009-0005-4530-9700

MODELLING PROCESSES OF UKRAINIAN-LANGUAGE TEXT LINGUISTIC ANALYSIS

Today, natural language text processing automation is widely used in various fields (public sector, science, education, media, everyday services, etc.). This requires appropriate software tools (services) that can provide such automated processing.

The purpose of this article is to analyse and study the problems of modelling the processes of linguistic analysis of Ukrainian natural language texts and designing appropriate software with the definition of the functional capabilities of its components.

The research methodology encompasses a comparative analysis of the primary software solutions in this subject area (linguistic analysis of texts presented in natural language), systematisation of approaches to automated text processing, and formalisation of these processes in the form of a corresponding process model.

The scientific novelty of the research lies in the analysis of current problems in systems supporting the processes of automated processing of texts in natural language, in particular Ukrainian; the development of a process model that combines different stages and phases of linguistic analysis of texts; and the design of an appropriate software solution to support this model.

Conclusions. The work examines the main problems of natural language text processing; identifies the main methods of processing Ukrainian-language texts; analyses and systematises modern systems that support individual stages of automated natural language text processing; a project for an author's software solution that will provide linguistic analysis of Ukrainian-language texts is described; a process model for linguistic analysis of natural language texts is proposed. The use of the proposed process model by users and developers will facilitate the deployment of high-quality Ukrainian-language services; on the part of institutions (relevant

ISSN 2617-796X (print)
ISSN 2618-0049 (online)

software products) during the modification of the linguistic analysis system, it will contribute to obtaining a more objective view of the most common problems of linguistic analysis (so-called language problems), lexicon gaps, and priorities for updates to Ukrainian-language NLP.

Keywords: natural language texts; linguistic analysis; modelling of linguistic analysis processes; process model; VESUM morphological dictionary; Universal Dependencies; disambiguation; tokenisation.

Надійшла 28.09.2025
Прийнята 04.11.2025
Стаття була вперше опублікована онлайн 29.12.2025

This is an open access journal, and all published articles are licensed under a Creative Commons Attribution 4.0.

ISSN 2617-796X (print)
ISSN 2618-0049 (online)