КОМП'ЮТЕРНІ НАУКИ

УДК 004.8-9 https://doi.org/10.32515/2664-262X.2026.13(44). 11-21

І.В. Козак, **В. А. Висоцька,** доц., д-р техн. наук, **Л.В. Чирун**, канд. техн. наук
Національний університет «Львівська політехніка», м. Львів, Україна
e-mail: ivan.v.kozak@lpnu.ua, victoria.a.vysotska@lpnu.ua, lyubomyr.v.chyrun@lpnu.ua

Інформаційна технологія квантитативного аналізу україномовного текстового контенту на основі DocBin-структур

У роботі представлено інформаційну технологію на основі розробки програмної підсистеми квантитативного аналізу українських текстів як компонента інформаційної системи обробки корпусних даних. Реалізовано алгоритми обчислення лексичних і морфологічних індексів (TTR, індекс Гоноре, індекс номінативності, частоти лем і POS) на основі структур DocBin із використанням spaCy та pandas. Проведено експериментальне тестування на корпусі з 7 документів обсягом близько 18 000 токенів. Час опрацювання 1 тис. токенів становив 0,11–6,72 с залежно від обраного NLP-агента. Результати підтвердили лінійну масштабованість, стабільність обчислень та можливість інтеграції підсистеми в захищені програмні середовища для аналітики текстових даних, зокрема в задачах моніторингу інформаційного простору та виявлення аномалій.

обробка природної мови, інформаційна система, аналіз даних, квантитативна лінгвістика, інформаційний моніторинг, корпусна лінгвістика, лексична різноманітність, TTR, індекс Гоноре, індекс номінативності, DocBin, spaCy, автоматизований аналіз тексту, українська мова

Постановка проблеми. Стрімке зростання обсягів текстових даних у цифровому середовищі формує потребу в ефективних засобах їх автоматизованого аналізу, структуризації та інтерпретації. Особливої актуальності це набуває для українськомовного сегмента інформаційного простору, де відсутність достатньо адаптованих інструментів кількісного лінгвістичного аналізу ускладнює проведення повноцінних досліджень і побудову прикладних систем, зокрема для квантитативного аналізу україномовних текстів. Створення адаптованої підсистеми квантитативного аналізу українських текстів є важливим кроком до формування національно орієнтованих інструментів цифрової лінгвістики та розвитку безпечних інтелектуальних інформаційних систем.

Аналіз останніх досліджень і публікацій. Аналіз лексичної різноманітності, індексу номінативності та частотних характеристик є ключовими аспектами корпусної лінгвістики, що дозволяють глибше зрозуміти структуру та функціонування мови [1]. Для української мови, яка є морфологічно багатою з одного боку та котра має менше прикладних засобів дослідження у порівнянні з іншими європейськими мовами, розробка та впровадження таких аналітичних інструментів є особливо актуальною.

Лексична різноманітність часто вимірюється за допомогою індексу типів до токенів $TTR=V/N$ (де V – кількість унікальних токенів чи текенів певного типу, N – загальна кількість токенів) та його модифікацій, таких як індекс Гоноре $G=V/\sqrt{N}$. Ці метрики дозволяють оцінити багатство словникового запасу в текстах різних жанрів та стилів. Однак, як зазначають в [2], існуючі інструменти для аналізу лексичної різноманітності здебільшого орієнтовані на англомовні корпуси, що створює труднощі при їх застосуванні до українських текстів. Індекс номінативності, який відображає

співвідношення іменників до інших частин мови в тексті, є важливим показником стилістичних та жанрових особливостей. Наприклад, наукові тексти зазвичай характеризуються високим рівнем номінативності. У роботі [3] підкреслюється необхідність розробки інструментів для автоматизованого обчислення цього індексу в українських текстах, оскільки існуючі рішення не забезпечують достатньої точності та адаптованості до морфологічних особливостей української мови.

Частотний (квантитативний) аналіз є фундаментальним методом у корпусній лінгвістиці, що дозволяє виявити найбільш уживані слова та конструкції в мові. Дослідження [1] демонструє застосування аналізу частот для оцінки ядра словника в українських текстах різних функціональних стилів. Вони використовують рангово-частотний аналіз для екстраполяції розміру словникового запасу, що є цінним підходом для побудови лексичних ресурсів. Крім того, інструмент StyloMetrix, спочатку розроблений для польської мови, був адаптований для української, що дозволяє аналізувати граматичні, стилістичні та синтаксичні патерни в текстах [4]. Цей інструмент використовує різноманітні метрики для класифікації текстів за стилем та жанром, що є корисним для дослідження лексичної різноманітності та інших характеристик [5-9]. Основними викликами при аналізі лексичної різноманітності, індексу номінативності та частотних характеристик в українських текстах є [1-5]:

– Морфологічна складність української мови: багатство флексій та словотворчих засобів ускладнює точну ідентифікацію лем та частин мови.

– Обмеженість існуючих інструментів: більшість доступних інструментів не адаптовані до специфіки української мови або мають обмежену функціональність.

– Відсутність стандартних корпусів: нестача великих, розмічених корпусів української мови ускладнює проведення статистичних аналізів.

Для подолання цих викликів необхідно розробляти нові інструменти, адаптовані до української мови, або модифікувати існуючі з урахуванням її особливостей [10-12]. Застосування формату збереження даних DocBin у SpaCy сприяє ефективному опрацюванні природної мови, в тому числі українськомовною конвенту [6-9].

Постановка завдання. Інформаційна технологія на основі комп’ютерної лінгвістики та Big Data аналізу базується на метриках, що відображають особливості лексики, граматики та стилю/жанру текстового контенту. Для української мови відсутні або обмежені доступні й адаптовані модульні інструменти, які б дозволяли автоматизовано обчислювати такі показники, як лексична різноманітність, індекс номінативності чи найпоширеніші частини мови. Це обмежує можливості дослідників у гуманітарних науках, лінгвістів, укладачів корпусів і розробників освітніх систем. Підсистема, що виконує обчислення таких характеристик на основі попередньо оброблених текстів, є актуальним і необхідним внеском у розвиток лінгвістичних прикладних рішень для української мови. Метою дослідження є розроблення методу для обчислення лексичних та граматичних індексів українських текстів на основі попередньо збережених у DocBin структурованих даних. Для досягнення вищезазначеної мети було поставлено наступні завдання:

1. Реалізувати механізм завантаження та опрацювання документів із DocBin.
2. Побудувати модулі для обчислення лексичної різноманітності.
3. Розрахунок індексу номінативності та визначення частот лем й частин мови.

Об’єкт дослідження – процеси аналізу лінгвальних характеристик текстів. Предмет дослідження – методи та засоби обчислення лексичних та морфологічних індексів (лексичної різноманітності, індексу номінативності, частот лем і частин мови) на основі структурованих лінгвальних даних. Запропоновано адаптацію підсистеми для обчислення лінгвістичних метрик на основі DocBin-структур із фокусом на українськомовні тексти. Отримано вдосконалене представлення мовних характеристик

тексту з урахуванням специфіки української мови. Запропонований підхід дозволяє масштабувати аналіз на інші мови, зберігаючи адаптивність і модульність архітектури.

Виклад основного матеріалу. Основною метою запропонованої інформаційної технології є забезпечення автоматизованого аналізу лінгвальних характеристик текстів (лексична різноманітність, частотні характеристики, індекс номінативності) на основі результатів морфологічної обробки. Це дозволить прикладним, корпусним та іншим лінгвістам отримувати кількісні оцінки мовної структури текстів без потреби в ручному аналізі. Аспекти генеральної мети:

1. Оцінка лексичної різноманітності – обчислення TTR, індексу Гоноре та інших метрик на основі лем.
2. Розрахунок морфологічних індексів, зокрема, індексу номінативності.
3. Частотний аналіз - виявлення найпоширеніших лем і частин мови у тексті.
4. Придатність до інтеграції дані повинні бути збереженими у вигляді, для легкого експорту чи візуалізації в подальшому.
5. Гнучкість метрик – можливість додавання нових індексів (напр. лексична щільність, ентропія).

Генеральна мета деталізується як набір інструментів, які мають покривати кілька площин: лексика, морфологія, частотний аналіз, адаптованість (рис. 1а). Виходячи з того, що функціонування підсистеми доцільне у комплексі з підсистемою попереднього опрацювання і розмітки текстів, варто також розглянути дерево цілей для всієї системи (рис. 1б). Критерії функціонування системи: підтримка обчислення метрик різноманітності, коректна класифікація частин мови для індексу номінативності та незалежність від зовнішнього API (працює офлайн). Альтернативні варіанти реалізації:

1. Власний Python-модуль, який працює з DocBin (вибраний варіант) – повний контроль, гнучкість, можливість адаптації під українську.
2. Інтеграція з існуючими бібліотеками (наприклад, stylo, lexdiv) – потребує адаптації до української мови, орієнтовані на англійську мову.
3. Використання генераторів корпусної статистики на кшталт Sketch Engine – закритість, обмеженість, не дає індексу номінативності чи TTR напряму.
4. Підрахунок вручну або в Excel – тривало, ненадійно, не масштабовано.

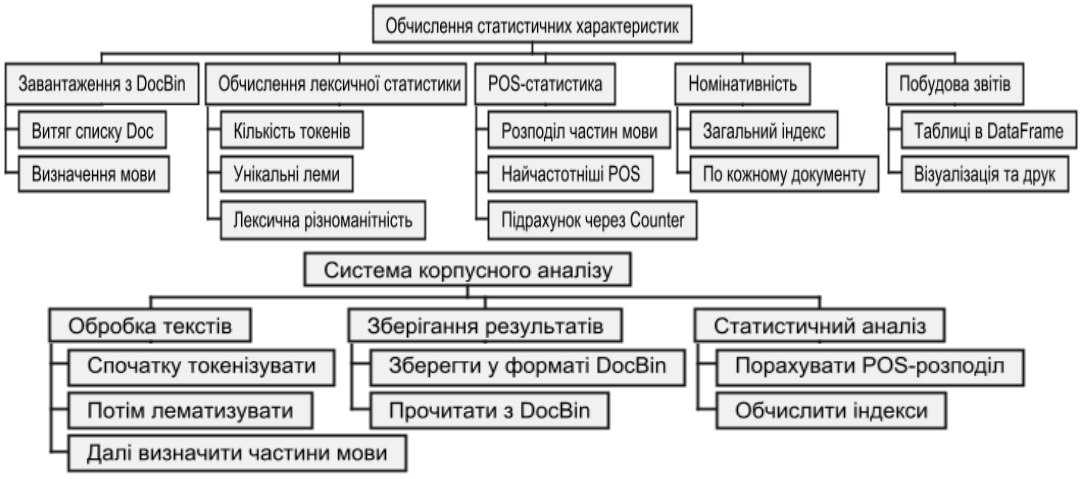


Рисунок 1 – Дерево цілей квантитативного/комплексного опрацювання текстів
Джерело: розроблено авторами

Підсистема виконує витяг та аналіз лінгвістичних характеристик тексту (рис. 2а) з файлів DocBin, отриманих від підсистеми попереднього опрацювання та розмітки текстів (загальний вигляд цілісної системи зображено на рис. 2б), зосереджуючись на оцінці стилістичних, морфологічних і частотних особливостей.

1. Завантаження DocBin (система імпортує збережені раніше оброблені документи з формату DocBin).

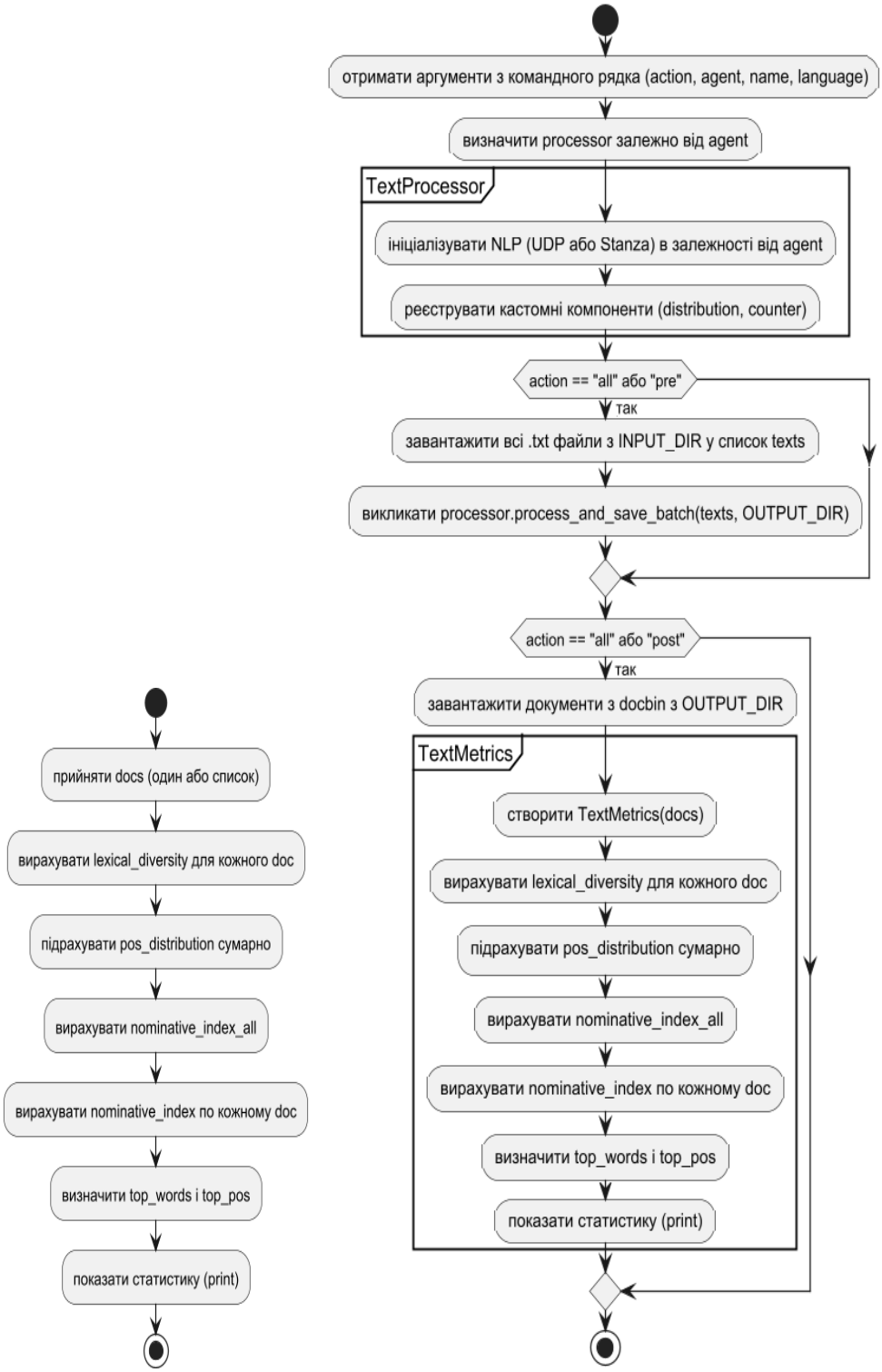


Рисунок 2 – Діаграма активностей для квантитативного опрацювання текстів та повної обробки текстів
Джерело: розроблено авторами

2. Витяг текстових та лінгвістичних атрибутів (для кожного документа отримуємо леми, POS-теги, частоти токенів).

3. Обчислення TTR (на основі кількості типів і токенів обчислюємо індекс лексичної різноманітності (TTR) й індекс Гоноре).

4. Розрахунок індексу номінативності (рахуємо співвідношення іменників до інших частин мови).

5. Агрегація частот (формуємо список найчастотніших лем і частин мови в кожному документі або у всьому корпусі).

6. Формування звіту/результатів (результати аналізу виводимо у GUI).

У діаграмі прецедентів для підсистеми квантитативного аналізу (рис. 3а) показано, як користувач (дослідник, лінгвіст або викладач) взаємодіє з підсистемою аналізу вже опрацьованих текстів. Основні сценарії включають: завантаження попередньо збереженого DocBin, запуск аналізу лексичної різноманітності (TTR, індекс Гоноре), розрахунок індексу номінативності (на основі POS-розмітки), отримання частот лем та частин мови, формування звіту з результатами аналізу. Ця діаграма демонструє, які саме функції система надає користувачеві для отримання кількісної лінгвальної інформації з розміченого корпусу текстів. Загальна діаграма прецедентів (рис. 4а) ілюструє, як користувач активує й використовує логіку усієї системи. Ця діаграма демонструє, які саме функції система надає користувачеві для отримання кількісної лінгвальної інформації з розміченого корпусу текстів.

Діаграма послідовностей у підсистемі (рис. 3б) моделює етапи виконання квантитативного аналізу одного документа. Вона описує: завантаження документа з DocBin, отримання необхідних атрибутів: лем, токенів, POS-тегів, запуск відповідних обчислювальних модулів (TTR, номінативність, тощо), створення агрегованого об’єкта зі статистикою, збереження / повернення результатів користувачеві. Порядок викликів, залежності між модулями та внутрішня логіка обчислень представлено в контексті часової взаємодії об’єктів загальної системи (рис. 4б).

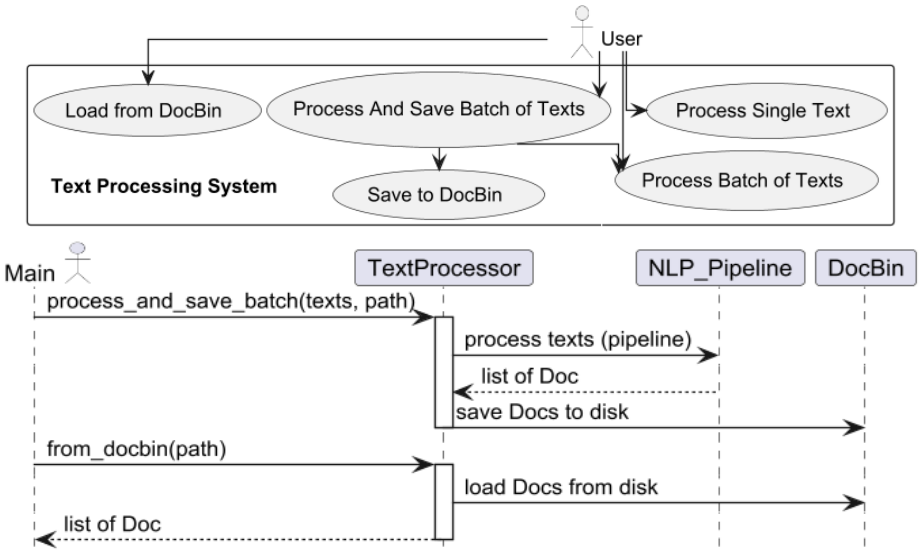


Рисунок 3 – Діаграма прецедентів/послідовностей квантитативного аналізу текстів
Джерело: розроблено авторами

Підсистема виконує розрахунок лінгвальних метрик (TTR, індекс номінативності), визначає найпоширеніші слова й частини мови. Вона слугує інструментом швидкого й надійного отримання квантитативної інформації у лінгвістичних дослідженнях. Застосовується в корпусній стилістиці, жанровому аналізі, типологічних дослідженнях, де важливо мати кількісні характеристики текстів. Також є цінним для освітніх платформ, де автоматизований аналіз дозволяє проводити практичні заняття з корпусної лінгвістики або розробляти словники чи граматики.

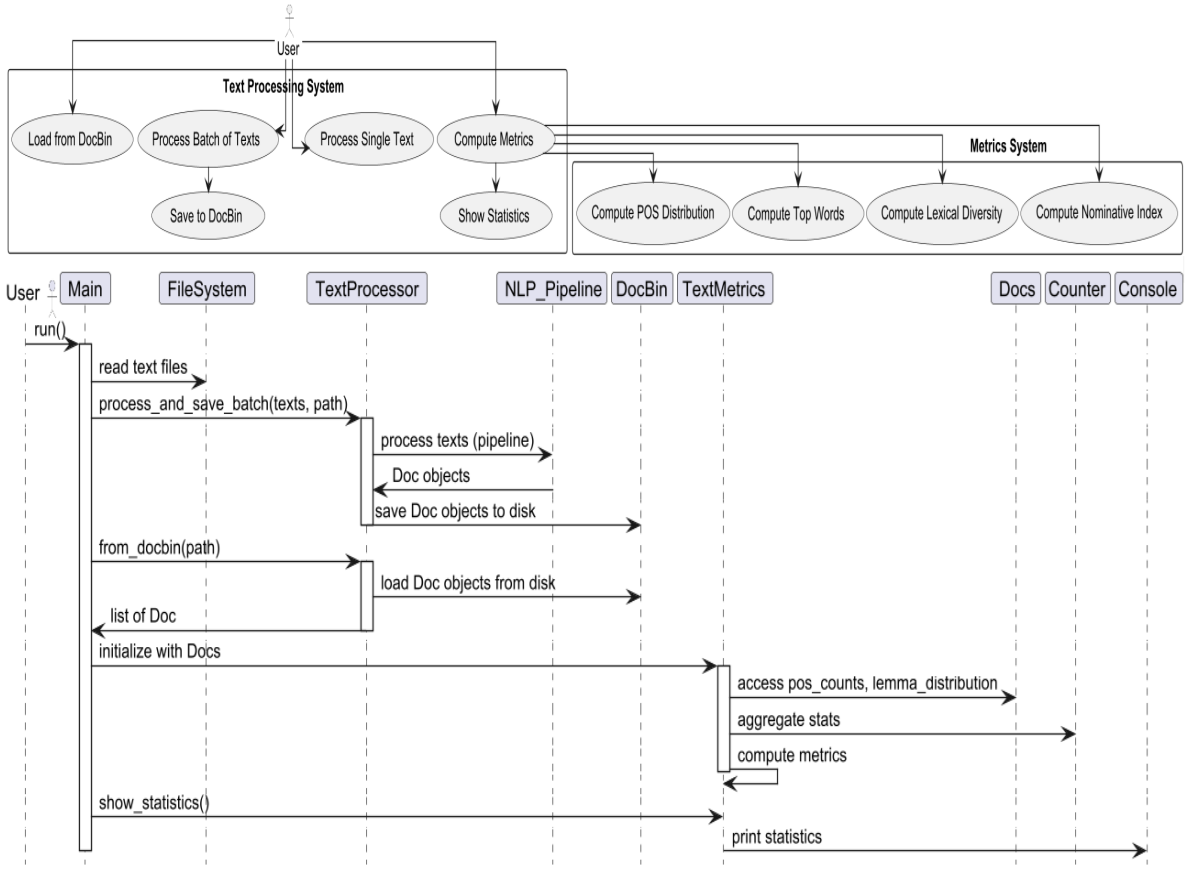


Рисунок 4 – Діаграма прецедентів/послідовностей цілісної системи
Джерело: розроблено авторами

У працях [7-10] відзначено запит на автоматичний аналіз корпусів, оскільки ручний аналіз займає багато часу і створює помилки в розрахунках. Крім того, більшість наявних інструментів не підтримують метрики лексичної різноманітності для української мови та не враховують морфологічної складності мови. Очікувані ефекти: швидке отримання індексів лексичної та морфологічної структури; підвищення об'єктивності аналізу текстів; покращення можливостей верифікації гіпотез у лінгвістичних дослідженнях. Вимоги до системи: підтримка індексів TTR, Гоноре, номінативності; агрегація частот POS і лем; вивід результатів у форматі, придатному для звіту чи візуалізації (DataFrame, CSV). Вхідні дані: DocBin, оброблені попередньою підсистемою. Вихідні: вивід для користувача результатів аналізу. Концептуальна модель: функції (витяг документів, підрахунок індексів, вивід статистики); структура (один клас (TextMetrics) з внутрішніми методами); модель (функціональна обробка документів у циклі з використанням pandas, Counter, та spaCy).

У другій підсистемі, яка відповідає за статистичний аналіз корпусу, основна увага зосереджена на обчисленні таких лінгвістичних метрик, як TTR, індекс Гоноре, індекс номінативності, частотність лем і частин мови. На відміну від першої підсистеми (паралельна підсистема, котра готує розмічені DocBin файли), де логічне виведення покладається на попередньо навчені моделі, тут використовуємо класичні числові алгоритми з відкритими формулами. Знання, на яких ґрунтується аналіз, зафіксовані у вигляді визначених метрик та способів їх обчислення. Наприклад, TTR визначається як відношення кількості унікальних лем до загальної кількості лем у тексті. Ми обрали для реалізації цієї підсистеми наступні бібліотеки як spaCy, collections.Counter та pandas. Підсистема зчитує документи у форматі DocBin, витягує з них леми й POS-теги,

обчислює статистичні дані (TTR, індекс номінативності, частотні леми), формує зведену статистику, яка може бути використана в подальшому для дослідження стилістики, складності або жанру тексту. Вхідними даними до підсистеми мають бути DocBin-файли, отримані з першої підсистеми. Підсистема має мати доступ до списку об'єктів Doc, словників частот (лем, POS) та датафрейм з результатами аналізу. Вихідні дані – статистичні таблиці (pandas.DataFrame), метрики (числові значення). Уся логіка реалізована у класі TextMetrics (рис. 5), який інкапсулює дані та аналітичні методи.

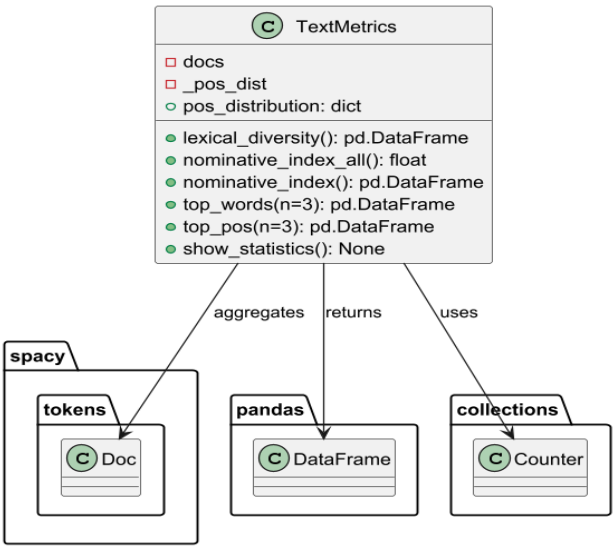


Рисунок 5 – Діаграма класів підсистеми квантитативного аналізу текстів.
Джерело: розроблено авторами

Основні функції (рис. 6):
– lexical_diversity() обчислює індекси TTR та Гоноре для кожного документа;
– nominative_index() визначає співвідношення іменників у кожному документі;
– nominative_index_all()розраховує індекс номінативності для всього корпусу;
– top_pos(n) виводить найчастотніші POS-теги;
– show_statistics() виводить статистику;
– top_words(n) виводить найчастотніші леми.

Підсистеми працюють послідовно, як два етапи одного конвеєру. Перша підсистема: читає текстові файли, проводить обробку, генерує DocBin. Друга підсистема: читає DocBin, аналізує розмічені документи, генерує звіти. Між ними не передається сира текстова інформація, а лише структуровані документи у форматі Doc, що гарантує консистентність лінгвістичних атрибутів (лем, POS, токенів) і дозволяє повторно використовувати корпус для різних цілей (стилометрія, жанровий аналіз, тощо). Ця інструкція користувача призначена для ознайомлення з функціональними можливостями інформаційної системи, розробленої для обробки та аналізу корпусів текстів українською мовою. Система складається з двох основних підсистем: попередньої обробки та статистичного аналізу. Керування програмою виконуємо з командного рядка передаючи параметри через аргументи. Основна команда виклику: *python main.py --action all --name corpora --agent stanza --language uk*. де: --action (-a) – визначає, яку частину запустити (pre – лише попередню обробку текстів (підсистема 1), post – лише аналітичну обробку (підсистема 2), all – повний цикл); --name (-n) – назва файлу або імені сесії, за яким будуть збережені DocBin і звіти.--agent – обробник (stanza або udpipe); --language – мова текстів (за замовчуванням uk).

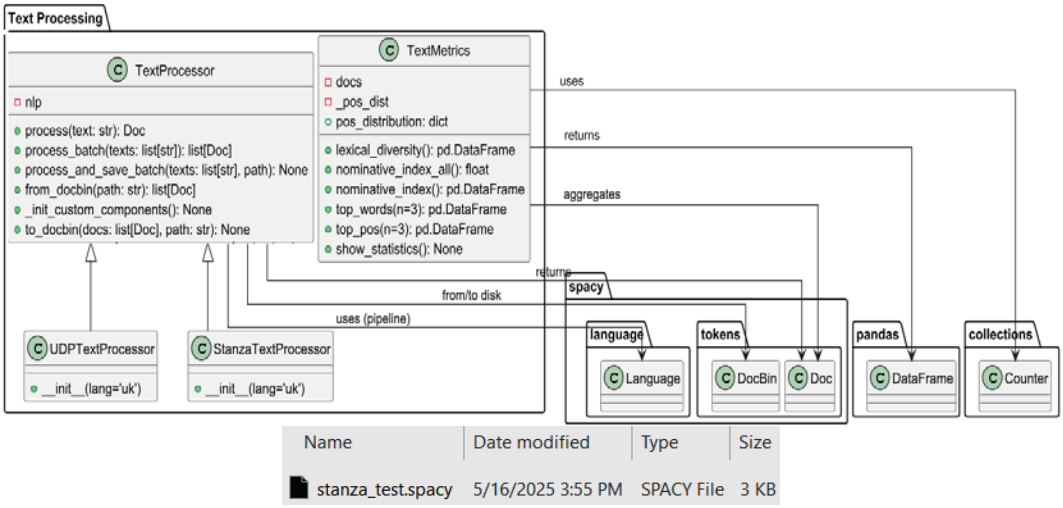


Рисунок 6 – Діаграма класів системи опрацювання текстів та Spacy DocBin файл
Джерело: розроблено авторами

Підсистема квантитативного опрацювання текстів виконує розрахунок індексу лексичної різноманітності (TTR), індексу номінативності, частот лем та частин мови. Використовує .spacy-файл, сформований попередньою підсистемою, з папки ./output/.

Приклад запуску лише аналізу (рис. 7а): *python main.py --action post --name test --agent stanza.* Результати аналізу буде виведено на екран користувача (рис. 7в).

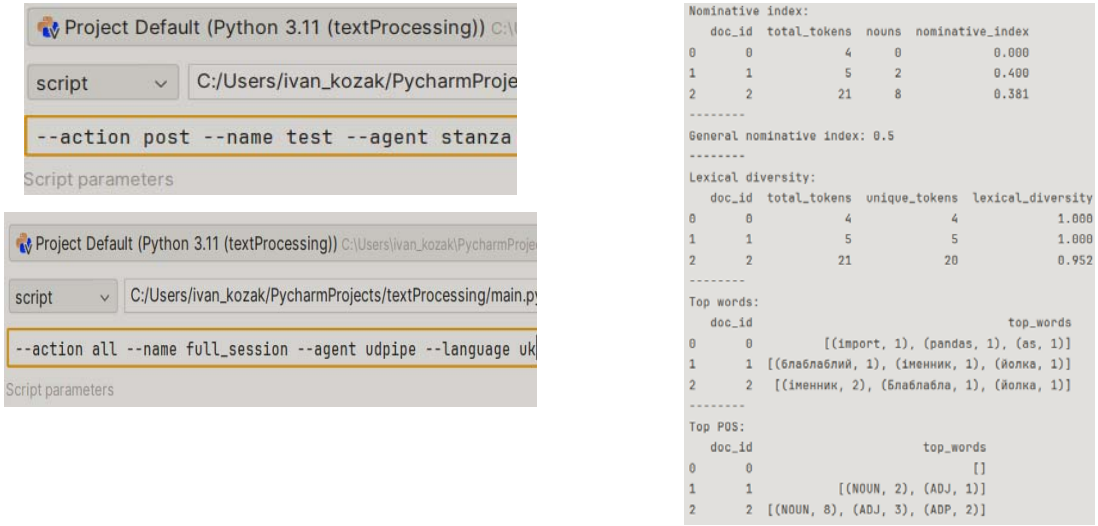
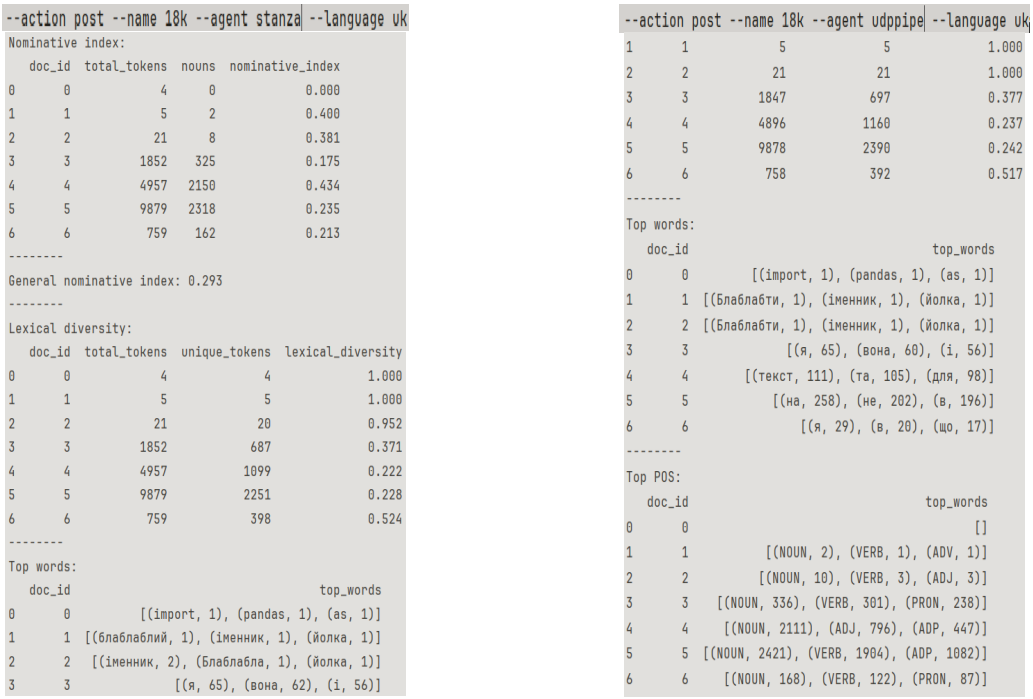


Рисунок 7 – Скріншот PyCharm конфігурації та результати опрацювання текстів
Джерело: розроблено авторами

Повна обробка (обидві підсистеми): *python main.py --action all --name full_session --agent udpipe --language uk.* Цей режим (рис. 7б) виконує повний цикл: від зчитування сирих текстів – до виведення статистичних даних. Системні вимоги: Python ≥ 3.9, установлені бібліотеки: spacy, spacy-stanza, spacy-udpipe, pandas, argparse та структура проєкту: */main.py → /definitions.py → /parser.py → /input/ → /output/.* Для перевірки роботи підсистеми квантитативної обробки текстів за допомогою агента spacy_stanza було проведено тестування на корпусі з 7 текстових файлів українською мовою загальним обсягом близько 18 000 токенів. Spacy DocBin файли розміщувалися у директорії ./output/. Запуск підсистеми здійснювався командою (рис. 8а): *python main.py --action post --name 18k --agent stanza --language uk.* Після виконання обробки результати аналізу було виведено в термінал користувача (рис. 8б).



```
--action post --name 18k --agent stanza --language uk
Nominative index:
   doc_id  total_tokens  nouns  nominative_index
0       0             4      0             0.000
1       1             5      2             0.400
2       2            21      8             0.381
3       3          1852    325             0.175
4       4          4957   2150             0.434
5       5          9879   2318             0.235
6       6           759    162             0.213
--------
General nominative index: 0.293
--------
Lexical diversity:
   doc_id  total_tokens  unique_tokens  lexical_diversity
0       0             4              4              1.000
1       1             5              5              1.000
2       2            21             20              0.952
3       3          1852            687              0.371
4       4          4957           1099              0.222
5       5          9879           2251              0.228
6       6           759            398              0.524
--------
Top words:
   doc_id                                  top_words
0       0          [(import, 1), (pandas, 1), (as, 1)]
1       1  [(блаблаблий, 1), (іменник, 1), (йолка, 1)]
2       2   [(іменник, 2), (блаблабла, 1), (йолка, 1)]
3       3              [(я, 65), (вона, 62), (і, 56)]
```

```
--action post --name 18k --agent udppipe --language uk
1       1            5             5             1.000
2       2           21            21             1.000
3       3         1847           697             0.377
4       4         4896          1160             0.237
5       5         9878          2390             0.242
6       6          758           392             0.517
--------
Top words:
   doc_id                                  top_words
0       0          [(import, 1), (pandas, 1), (as, 1)]
1       1  [(Блаблабти, 1), (іменник, 1), (йолка, 1)]
2       2  [(Блаблабти, 1), (іменник, 1), (йолка, 1)]
3       3              [(я, 65), (вона, 60), (і, 56)]
4       4       [(текст, 111), (та, 105), (для, 98)]
5       5           [(на, 258), (не, 202), (в, 196)]
6       6              [(я, 29), (в, 20), (що, 17)]
--------
Top POS:
   doc_id                                  top_words
0       0                                         []
1       1          [(NOUN, 2), (VERB, 1), (ADV, 1)]
2       2         [(NOUN, 10), (VERB, 3), (ADJ, 3)]
3       3   [(NOUN, 336), (VERB, 301), (PRON, 238)]
4       4    [(NOUN, 2111), (ADJ, 796), (ADP, 447)]
5       5  [(NOUN, 2421), (VERB, 1904), (ADP, 1082)]
6       6    [(NOUN, 168), (VERB, 122), (PRON, 87)]
```

Рисунок 8 – Результат квантитативного аналізу текстів згенерований на основі stanza/ udppipe
Джерело: розроблено авторами

Аналогічним способом запуск підсистеми для агента udppipe можна здійснити командою (рис. 8в): *python main.py --action post --name 18k --agent udppipe --language uk.* Після виконання обробки результати аналізу було виведено в термінал користувача (рис. 8г). Аналіз часу виконання було виконано для обох агентів на корпусах текстів з прогресивно зростаючою кількістю токенів (1, 3, 8 і 18 тис. токенів) (див. табл. 1). Як видно, час виконання підсистеми аналізу частково залежить від обраного агента, оскільки більшість часу опрацювання файлів за допомогою spacy_stanza займає завантаження та ініціалізація його процесорів.

Таблиця 1 – Час опрацювання текстів підсистемою

Кількість токенів	Кількість документів	Час виконання (сек)		Час опрацювання тисячі токенів (сек)	
		stanza	udppipe	stanza	udppipe
1000	4	6,72	1,31	6,72	1,31
3000	5	7,09	1,53	2,36	0,51
8000	6	6,89	1,46	0,86	0,18
18000	7	8,9	1,97	0,49	0,11

Джерело: розроблено авторами

Водночас час виконання залишається стабільним при різних обсягах даних і лише дещо залежить від кількості опрацьованих файлів корпусу (але не їх розміру), оскільки більшість роботи було виконано на етапі попереднього опрацювання текстів. Це свідчить про універсальність аналітичного модуля.

Висновки. Розроблена інформаційна технологія аналізу лінгвістичних характеристик тексту є ефективним інструментом для автоматичного обчислення ключових лінгвістичних метрик, таких як TTR (Type-Token Ratio), індекс Гоноре, індекс номінативності, а також для визначення частот найпоширеніших лем та частин мови у корпусі текстів. Основною метою цієї підсистеми було створення гнучкого й масштабованого засобу для отримання кількісних характеристик текстів українською мовою, що дозволяє проводити статистичний аналіз корпусів будь-якого розміру. Результати експериментальної апробації сумісність підсистеми із DocBin-структурами. Аналіз результатів показав, що система правильно обробляє дані незалежно від того, який агент (модель) використовувався на етапі попередньої обробки (spacy_stanza чи

spacy_udpipe), а час виконання зростає лінійно зі збільшенням кількості текстів корпусу. Це свідчить про добру масштабованість рішення та підтверджує його практичну придатність для обробки великих обсягів текстів.

Перспективи розвитку ІТ квантитативного аналізу україномовного текстового контенту на основі DocBin-структур пов'язані із збільшенням множини нерозв'язаних задач опрацювання природньої мови на основі списку лінгвістичних метрик. В майбутніх дослідженнях ми плануємо доповнити конвеєр квантитативного аналізу україномовного тексту аналізом показників складності тексту та синтаксичних конструкцій, а також виявлення колекції рідковживаних слів. Отже, підсистема аналізу лінгвістичних характеристик тексту є надійним та гнучким інструментом для кількісного аналізу корпусів. Її подальший розвиток дозволить розширити спектр досліджень, зробити аналіз більш глибоким та багатовимірним, а також адаптувати систему до різноманітних дослідницьких завдань у сфері цифрової лінгвістики.

Список літератури

1. Buk S. N., Rovenchak A. A. The Rank-Frequency Analysis for the Functional Style Corpora in the Ukrainian Language. Journal of Quantitative Linguistics. 2004. Vol. 11, No. 3. P. 161–171. DOI: https://doi.org/10.1080/0929617042000314912.
2. Козак І., Кунанець Н. Проблеми та виклики при створенні корпусу українських текстів. Науковий вісник НУЛП. 2023. № 4. С. 101–108. DOI: https://doi.org/10.36930/40340213.
3. Kozak I., Kunanets N. Information systems for working with text corpora: classification and comparative analysis. Вісник «Інформаційні системи та мережі» Національного університету «Львівська політехніка». 2024. Вип. 16. С. 273–289. DOI: https://doi.org/10.23939/sisn2024.16.273.
4. Stetsenko D., Okulska I. The Grammar and Syntax Based Corpus Analysis Tool For The Ukrainian Language. Communication Papers of the 18th Conference on Computer Science and Intelligence Systems. 2023. P. 303–311. DOI: https://doi.org/10.48550/arXiv.2305.13530.
5. Козак І., Кунанець Н. Information system for text corpora management through the lens of business requirements. Інформаційні технології: теорія і практика : тези доповідей II (VIII) Міжнар. наук.-практ. конф. здобувачів вищої освіти і молодих учених ІТТП-2025 (Запоріжжя, 2025). Запоріжжя : НУ «Запорізька політехніка», 2025. С. 55–58.
6. Chiarcos C. CoNLL-Merge: Efficient harmonization of concurrent tokenization and textual variation. Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI). Association for Computational Linguistics, 2021. P. 41–52.
7. Explosion AI. spaCy documentation: DocBin serialization. 2023. URL: https://spacy.io/api/docbin (дата звернення: 18.02.2026).
8. Федчук Р., Висоцька В. Інформаційні технології вирішення задачі виправлення помилок в україномовних текстах. Вісник «Інформаційні системи та мережі» Національного університету «Львівська політехніка». 2024. Вип. 16. С. 11–34. URL: https://doi.org/10.23939/sisn2024.16.011.
9. Хоптяр А. О., Катуніна О. С., Калініченко Т. М. Використання мовних корпусів у дослідженні усного та письмового перекладу: Аналіз великомасштабних мовних даних. Вісник науки та освіти. Серія: Філологія. 2024. № 9(27). С. 500–514. DOI: https://doi.org/10.52058/2786-6165-2024-9(27)-500-514.
10. Anthony L. AntConc (Version 3.4.4) [Computer software]. Waseda University, 2013. URL: https://www.laurenceanthony.net/software/antconc (дата звернення: 18.02.2026).
11. Kilgarriff A., Baisa V., Bušta J. et al. The Sketch Engine. Lexicography. 2014. Vol. 1, № 1. P. 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9.
12. Straka M., Hajic J., Straková J. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, 2016. P. 4290–4297.

References

1. Buk, S. N., & Rovenchak, A. A. (2004). The Rank-Frequency Analysis for the Functional Style Corpora in the Ukrainian Language. Journal of Quantitative Linguistics, 11(3), 161–171. https://doi.org/10.1080/0929617042000314912
2. Kozak, I., & Kunanets, N. (2023). Problems and challenges in creating a corpus of Ukrainian texts. Naukovyi visnyk NULP, (4), 101–108 [in Ukrainian]. https://doi.org/10.36930/40340213
3. Kozak, I., & Kunanets, N. (2024). Information systems for working with text corpora: classification and comparative analysis. Visnyk «Informatsiini systemy ta merezhi» Natsionalnoho universytetu «Lvivska politekhnika», (2), 273–289. https://doi.org/10.23939/sisn2024.16.273

4. Stetsenko, D., & Okulska, I. (2023). The Grammar and Syntax Based Corpus Analysis Tool For The Ukrainian Language. Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, 303–311. https://doi.org/10.48550/arXiv.2305.13530
5. Kozak, I., & Kunanets, N. (2025). Information system for text corpora management through the lens of business requirements. Informatsiini tekhnolohii: teoriia i praktyka: materialy II (VIII) Mizhnarodnoi naukovo-praktychnoi konferentsii (pp. 55–58). Zaporizhzhia: NU «Zaporizka politekhnika» [in Ukrainian].
6. Chiarcos, C. (2021). CoNLL-Merge: Efficient harmonization of concurrent tokenization and textual variation. Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) (pp. 41–52). Association for Computational Linguistics.
7. Explosion AI. (2023). spaCy documentation: DocBin serialization. https://spacy.io/api/docbin
8. Fedchuk, R., & Vysotska, V. (2024). Information technologies for solving the problem of error correction in Ukrainian-language texts. Visnyk «Informatsiini systemy ta merezhi» Natsionalnoho universytetu «Lvivska politekhnika», 16, 11–34 [in Ukrainian]. https://doi.org/10.23939/sisn2024.16.011
9. Khoptyar, A. O., Katunina, O. S., & Kalinichenko, T. M. (2024). The use of language corpora in the study of oral and written translation: Analysis of large-scale language data. Visnyk nauky ta osvity. Seriia: Filolohiia, 9(27), 500–514 [in Ukrainian]. https://doi.org/10.52058/2786-6165-2024-9(27)-500-514
10. Anthony, L. (2013). AntConc (Version 3.4.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc
11. Kilgarriff, A., Baisa, V., Bušta, J., et al. (2014). The Sketch Engine. Lexicography, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9
12. Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016) (pp. 4290–4297). European Language Resources Association.

Ivan Kozak, Victoria Vysotska, Assoc. Prof., Doct. tech. sci., **Lyubomyr Chyrun,** PhD tech. sci
Lviv Polytechnic National University, Lviv, Ukraine

Information Technology for Quantitative Analysis of Ukrainian-Language Textual Content Based on DocBin Structures

The purpose of this article is to develop and implement information technology for quantitative text analysis for Ukrainian-language corpora within the framework of a modular information processing system. The study aims to develop algorithmic tools for automatic computation of lexical and morphological indices based on structured linguistic data. Particular attention is paid to ensuring scalability, offline functionality, and compatibility with secure computing environments. The proposed solution is oriented toward applications in computer science, data analytics, and cybersecurity, where reliable and reproducible text metrics are required. The research also seeks to address the lack of adapted quantitative tools for morphologically rich languages such as Ukrainian.

The implemented subsystem operates on preprocessed DocBin structures and performs automated extraction of lemmas, tokens, and part-of-speech tags using spaCy-based pipelines. Algorithms for calculating Type-Token Ratio (TTR), Honore's index, nominative index, and frequency distributions of lemmas and POS tags were developed and integrated into a unified TextMetrics class. The architecture follows a modular design that separates preprocessing and statistical analysis stages, ensuring extensibility and maintainability. Experimental validation was conducted on a corpus of 7 Ukrainian texts with a total volume of approximately 18,000 tokens. Performance evaluation demonstrated stable execution time and linear scalability with respect to corpus size. Processing time per 1,000 tokens ranged from 0.11 to 6.72 seconds depending on the selected NLP agent. The subsystem produces structured statistical outputs in tabular formats suitable for further visualization, reporting, or integration into analytical platforms. The design supports deployment in offline environments, reducing risks related to data leakage and enhancing applicability in protected infrastructures.

The results confirm the correctness, robustness, and scalability of the developed subsystem for quantitative linguistic analysis. The approach enables efficient extraction of measurable textual characteristics and can be integrated into broader information systems for corpus management, anomaly detection, and information monitoring. The proposed solution contributes to the advancement of computational linguistics tools for Ukrainian and supports interdisciplinary applications in computer science and cybersecurity. Future development includes performance optimization for large-scale corpora and extension of the metric set with syntactic and complexity-based indicators.

natural language processing, information system, data analysis, quantitative linguistics, information monitoring, corpus linguistics, lexical diversity, TTR, Honoré index, nominative index, DocBin, spaCy, automated text analysis, Ukrainian language

Одержано (Received) 18.12.2025 *Прорецензовано (Reviewed) 20.02.2026*
Прийнято до друку (Approved) 24.02.2026