УДК 81'322:81'25:004.9
DOI https://doi.org/10.32782/tps2663-4880/2026.49.45

МЕТОДОЛОГІЯ ДОПЕРЕКЛАДАЦЬКОГО АНАЛІЗУ ТЕКСТУ ІЗ ЗАСТОСУВАННЯМ ІНСТРУМЕНТІВ NLP

METHODOLOGY FOR PRE-TRANSLATION TEXT ANALYSIS USING NLP TOOLS

Рєзіна О.В.,
orcid.org/0000-0001-6077-9413
кандидат педагогічних наук, доцент,
доцент кафедри інформаційних та цифрових технологій
Центральноукраїнського державного університету імені Володимира Винниченка

Ярова Л.О.,
orcid.org/0000-0001-6817-1787
кандидат педагогічних наук,
вчитель англійської мови
Комунального закладу «Ліцей «Науковий Кропивницької міської ради»»

У статті обґрунтовано можливості застосування інструментів обробки природної мови (Natural Language Processing, NLP) для допереклададацького аналізу англомовних текстів і формування оптимальної перекладацької стратегії. Актуальність дослідження зумовлена недостатньою розробленістю питання інтегрування NLP-алгоритмів у підготовчий етап перекладу. За теоретичну основу обрано модель К. Норд, що передбачає розмежування позатекстових і внутрішньотекстових факторів. Матеріалом дослідження стали три англомовні тексти спільної тематики штучного інтелекту, що репрезентують різні жанри: науково-технічний, науково-популярний та рекламний. Для аналізу використано Python-бібліотеку spaCy. Дослідження зосереджено на трьох параметрах: лексичних характеристиках, іменованих сутностях (Named Entity Recognition, NER) і синтаксичній структурі речень. Проаналізовано частотний розподіл частин мови, лексичну щільність, кількість і типи іменованих сутностей, середню та максимальну довжину речень та глибину синтаксичних дерев. Результати засвідчили виразні жанрові відмінності. Науково-технічний текст характеризується найвищою синтаксичною складністю та лексичною щільністю, науково-популярний – більшою часткою прислівників, коротшими реченнями та ширшим соціально-географічним контекстом, а рекламний – переважанням іменників, а також насиченістю назвами брендів і продуктів. Виявлено, що близька лексична щільність різножанрових текстів не означає однакової складності їх перекладу. Отримані показники дають змогу диференціювати стратегії: забезпечувати термінологічну точність у науковому тексті, зберігати доступність та емоційно-оцінну забарвленість науково-популярного дискурсу, а в рекламному – відтворювати назви продуктів, брендів і рекламний ефект. Доведено, що поєднання моделі Christiane Nord із можливостями spaCy підвищує об'єктивність і системність допереклададацького аналізу та забезпечує перекладача емпіричними даними до початку перекладу.

Ключові слова: обробка природної мови, spaCy, допереклададацький аналіз перекладацька стратегія, лексична щільність, іменовані сутності, синтаксичний аналіз, штучний інтелект.

This article explores the potential of using natural language processing (NLP) tools for pre-translation analysis of English-language texts and for developing an optimal translation strategy. The relevance of this study stems from the fact that the integration of NLP algorithms into the preparatory stage of translation has not been sufficiently explored. The theoretical framework is based on Christiane Nord's model, which distinguishes between extratextual and intratextual factors. The research material consisted of three English-language texts on the common topic of artificial intelligence, representing different genres: scientific and technical, popular science, and advertising. The spaCy Python library was used for the analysis. The study focuses on three parameters: lexical characteristics, named entities (Named Entity Recognition, NER), and the syntactic structure of sentences. We analyzed the frequency distribution of parts of speech, lexical density, the number and types of named entities, the average and maximum sentence lengths, and the depth of syntactic trees. The results revealed distinct genre differences. Scientific and technical texts are characterized by the highest syntactic complexity and lexical density; popular science texts – by a higher proportion of adverbs, shorter sentences, and a broader socio-geographical context; and advertising texts – by a predominance of nouns, as well as an abundance of brand and product names. It has been found that a similar lexical density across texts of different genres does not imply that they are equally difficult to translate. The obtained indicators make it possible to differentiate strategies: to ensure terminological accuracy in scientific texts, to preserve accessibility and the emotional and evaluative tone of popular science discourse, and in advertising texts – to reproduce product and brand names and the advertising effect. It has been demonstrated that combining Christiane Nord's model with the capabilities of spaCy enhances the objectivity and systematic nature of pre-translation analysis and provides the translator with empirical data before beginning the translation.

Key words: natural language processing, spaCy, translation strategy, pre-translation analysis, lexical density, named entities, syntactic analysis, artificial intelligence.

Постановка проблеми. Сучасні інструменти обробки природної мови (Natural Language Processing, NLP) кардинально змінили підходи до автоматизованого лінгвістичного аналізу тексту, їхньою ключовою перевагою є об'єктивність і точність морфологічного, синтаксичного та семантичного аналізу. Зокрема Python-бібліотека spaCy забезпечує широкий спектр лінгвістичних функцій – від визначення частин мови та лематизації до синтаксичного аналізу на основі дерев залежностей і розпізнавання іменованих сутностей.

Попри значний потенціал NLP-інструментів, їхнє застосування у доперекладацькому аналізі тексту залишається малодослідженим: більшість наявних розвідок зосереджена на машинному перекладі, тоді як підготовчий етап перекладацького процесу, аналіз вихідного тексту, не отримав належної уваги в контексті NLP. Актуальність цього дослідження зумовлена необхідністю розроблення методології застосування бібліотеки spaCy як інструменту доперекладацького аналізу, здатного забезпечити перекладача об'єктивними кількісними даними про лексичні та синтаксичні характеристики вихідного тексту.

Метою статті є обґрунтування можливостей використання Python-бібліотеки spaCy у доперекладацькому аналізі англомовних текстів різних жанрів для формування оптимальної перекладацької стратегії.

Аналіз останніх досліджень і публікацій. Доперекладацький аналіз тексту – це комплексна процедура дослідження вихідного тексту, спрямована на виявлення його комунікативної мети та лінгвістичних характеристик для формування оптимальної стратегії перекладу.

К Сересова та Д. Бревенікова вважають, що аналіз тексту – це вирішальний процес, що передує перекладу, і полягає у дослідженні внутрішніх (лінгвістичних) та зовнішніх (контекстуальних) факторів тексту оригіналу для забезпечення точності, функціональності та культурної відповідності цільового тексту. Такий аналіз дає можливість перекладачам зрозуміти призначене повідомлення, функцію та цільову аудиторію, прийняти обґрунтовані рішення щодо стилю, термінології та стратегії [12].

У своїй фундаментальній праці К. Норд [9] представляє аналіз вихідного тексту як крок, на якому фактори комунікативної ситуації є надзвичайно важливими, оскільки вони визначають комунікативну функцію перекладу, яка вважається «вирішальним критерієм для текстуально підпорядкованих семантичних та синтаксичних особливостей вихідного тексту». Фактори комунікативної ситуації дослідниця поділяє на дві категорії: позатекстові та внутрішньотекстові чинники аналізу. До першої категорії відносяться: автор, його намір, цільова аудиторія, засіб/канал передавання тексту, місце та час комунікації, мотив комунікації та функція тексту. Тоді як друга включає тематику, зміст, пресупозицію, композицію тексту, невербальні елементи, лексику, структуру речення та надсегментні ознаки.

Ідеї функціонального аналізу К. Норд знаходять свій розвиток у сучасних працях вітчизняних науковців. Зокрема, у дослідженні [10] продовжується вивчення комунікативних факторів перекладу. У роботі [5] автентичний текст досліджується за допомогою функцій та елементів моделі К. Норд. Тут розкриваються такі функціональні характеристики як зміст та структура вихідного тексту і доводиться, що зазначена модель може бути застосована в перекладознавстві до різних типів текстів. Основна увага статті Г. Гаспарян [4] сфокусована на взаємозв'язку та взаємодії таких видів аналізу вихідного тексту як прагматичний, дискурсивний та доперекладацький, що розглядаються як суттєві вимоги для релевантного перекладу.

Поряд із загальним аналізом, окремої уваги набуває етап безпосереднього доперекладацького читання. Робота [2] має на меті виявити внесок, який доперекладацьке читання може надати перекладачеві під час процесу перекладу.

У своїй оглядовій статті Р. Міята [8] систематизує 57 термінів, що позначають властивості вихідного документа, і розподіляє їх за чотирма основними категоріями: знання, комунікація, структура та текстові властивості.

З огляду на вищесказане, можна стверджувати, що аналіз вихідного тексту є першим і важливим кроком до успішного перекладу. Деякі його етапи підлягають автоматизації засобами NLP. Питанням автоматичної обробки тексту присвячено низку робіт. У ґрунтовному методологічному дослідженні [3] розглядається метод некерованого тематичного моделювання з використанням структурної тематичної моделі. Ця модель дає можливість аналізувати метадані (автор, країна походження та час написання тексту) для глибшого розуміння зв'язків між цими даними та темами, які представлені в корпусі текстів. Обговорюються особливості попередньої обробки тексту: видалення стоп-слів, стемінг та лематизація, обробка складних слів, проблеми сегментації для різних мов.

А. Мацанович досліджує інструменти аналізу тексту для соціальних наук. Автор аналізує

сучасні дослідження та показує, як словникові методи, інструменти семантичного та мережевого аналізу, мовні моделі, машинне навчання можуть допомогти соціологам у різних аналітичних завданнях. Також у цій роботі демонструється кілька важливих наслідків використання великих наборів даних та обчислювальних методів для визначення складного змісту в текстах [7].

Дослідження гендерного розриву серед студентів, що навчаються на освітніх програмах з інформаційних технологій та інженерії Міжнародного університету Парагон (ParagonIU), було проведено з використанням текстового аналізу. Попередня обробка неструктурованих текстів, які були відповідями з анкети опитування, здійснювалася засобами мови програмування Python із використанням бібліотек Natural Language Toolkit (NLTK) та NumPy. Застосовані інструменти забезпечили достатній рівень надійності та валідності отриманих результатів [11].

Отже, доперекладацький аналіз вихідного тексту є фундаментальним етапом, який поєднує дослідження позатекстового контексту та мовних особливостей оригіналу. Проте у сучасній літературі залишається недостатньо висвітленим питання комплексного інтегрування NLP-алгоритмів безпосередньо у практику доперекладацького аналізу. Тому вивчення того, які саме внутрішньотекстові фактори моделі К. Норд можна кількісно виміряти за допомогою сучасних NLP-інструментів і як ці вимірювання можуть бути інтерпретовані для формування перекладацької стратегії є необхідним.

Виклад основного матеріалу. Відповідно до моделі К. Норд, доперекладацький аналіз доцільно розширити методами комп'ютерної лінгвістики. Якщо позатекстові фактори (намір, аудиторія, мотив комунікації) потребують інтерпретації людиною, то внутрішньотекстові параметри (лексика, тематика, структура речення, надсегментні ознаки) можуть бути виміряні за допомогою NLP-інструментів. Наприклад, використання методів автоматизованого видобування термінів дозволяє перекладачеві швидко ідентифікувати складні лексеми, що значно пришвидшує етап доперекладацького читання, описаний М. Байраком [2].

У межах цієї роботи автоматизований аналіз зосереджено на *трьох ключових лінгвістичних параметрах*, що мають найбільше практичне значення для доперекладацького аналізу:

1. лексичних характеристиках (частотний розподіл частин мови та лексична щільність);

2. іменованих сутностях (Named Entity Recognition, NER);

3. синтаксичній структурі (довжина речень, глибина синтаксичних залежностей).

Ці три параметри забезпечують достатню емпіричну базу для формування диференційованих перекладацьких стратегій.

Для реалізації доперекладацького аналізу внутрішньотекстових факторів, *як інструмент обробки текстів*, було обрано Python-бібліотеку spaCy. Вона забезпечує лінгвістично обґрунтовану токенізацію, містить компоненти для розпізнавання іменованих сутностей, визначення частин мови, аналізу залежностей, сегментації речень, класифікації тексту, лематизації, морфологічного аналізу, зв'язування сутностей та інше (https://spacy.io/). При визначенні можливостей покращення автоматизованого розрахунку показників лексичного багатства для текстів англійською мовою SpaCy надав показники, які найбільше корелювали з людською оцінкою, у порівнянні з аналізатором лексичної складності (Lexical Complexity Analyzer, LCA) та NLTK [15]. Крім того, для забезпечення вищої точності результатів було застосовано мовну модель en_core_web_trf, яка використовує модель трансформатора на основі RoBERTa для забезпечення розширених лінгвістичних анотацій, таких як тегування частин мови, парсинг залежностей та розпізнавання іменованих сутностей (https://spacy.io/models/en).

Для апробації обраного інструментарію та проведення NLP-аналізу як матеріал дослідження було обрано *три англомовні тексти тематики штучного інтелекту (ШІ)*, що належать до різних жанрів і є у відкритому доступі. Обсяг кожного тексту становить від 1100 до 1250 слів, що забезпечує їх порівнянність для подальшого NLP-аналізу.

Текст А є науково-технічною статтею Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement (Yin et al., 2024), опублікованою на платформі arXiv (URL: https://arxiv.org/abs/2410.04444). Для аналізу було відібрано три структурні компоненти статті: Анотація, Вступ та Висновки, що у сукупності становить приблизно 1100 слів. Обрані розділи є семантично самодостатніми і репрезентують жанрові характеристики науково-технічного тексту повною мірою: Анотація формулює наукову проблему та результати, Вступ окреслює теоретичне підґрунтя, Висновки підсумовують здобутки та визначають перспективи.

Текст В – науково-популярна стаття Is Recursive Self-Improvement the Dawning of AI Superintelligence?, опублікована на платформі The

Conversation (URL: https://theconversation.com/is-recursive-self-improvement-the-dawning-of-ai-superintelligence-287081). Вона розрахована на широку аудиторію та висвітлює концепцію рекурсивного самовдосконалення ШІ у доступній формі. Обсяг тексту становить приблизно 1200 слів.

Текст С сформовано з двох рекламних прес-релізів компаній б'юті-індустрії, що просувають ШІ-продукти: Perfect Corp. Unveils Next-Generation AI Beauty Agent and API Innovations Transforming Beauty, Skincare, and Retail at CES 2026 (URL: https://www.businesswire.com/news/home/20251219606970/en/Perfect-Corp.-Unveils-Next-Generation-AI-Beauty-Agent-and-API-Innovations-Transforming-Beauty-Skincare-and-Retail-at-CES-2026) та Haut.AI Launches First Consumer Version of Skin.Chat, Its Beauty AI Agent (URL: https://www.prnewswire.com/news-releases/hautai-launches-first-consumer-version-of-skinchat-its-beauty-ai-agent--making-holiday-beauty-shopping-on-amazon-seamless-302621110.html), опублікованих відповідно на платформах Business Wire та PR Newswire. Обсяг тексту становить приблизно 1200 слів.

Усі три тексти об'єднані спільною темою – штучним інтелектом, що унеможливлює вплив тематичних відмінностей на результати NLP-аналізу. У процесі підготовчого етапу для підвищення точності результатів із тексту А були вилучені бібліографічні посилання у форматі (Автор, рік), наприклад, (Bommasani et al., 2021; Bubeck et al., 2023; Zhang et al., 2024).

Після попередньої обробки та стандартизації текстового корпусу було здійснено аналіз його лексичних характеристик, що розпочався зі встановлення частотного розподілу частин мови.

Відповідно до положень корпусної лінгвістики, частотний розподіл частин мови є значущим стилістичним індикатором, що відображає жанрову належність тексту. Дослідниками встановлено, що висока питома вага іменників корелює з інформаційно орієнтованим академічним стилем, тоді як переважання дієслів та прислівників характерне для більш розмовних жанрів [13].

На рисунку 1 представлено частотний розподіл частин мови для досліджуваних текстів.

Аналіз розподілу іменників виявив найвищу їхню частку у тексті С (57.20%), що пояснюється жанровою специфікою рекламного прес-релізу: він насичений назвами продуктів, технологій та функцій (*AI Beauty Agent*, *skin analysis*, *virtual try-on*, *API Innovation Suite*), які утворюють щільні іменникові групи. Тексти А та В демонструють близькі показники (50.47% та 46.31%), що свідчить про схожу іменникову насиченість академічного та науково-популярного жанру і узгоджується з даними корпусних досліджень щодо номінативного стилю наукового письма [13].

Розподіл дієслів демонструє чітку спадну тенденцію: А (25.16%) → В (24.20%) → С (23.10%).

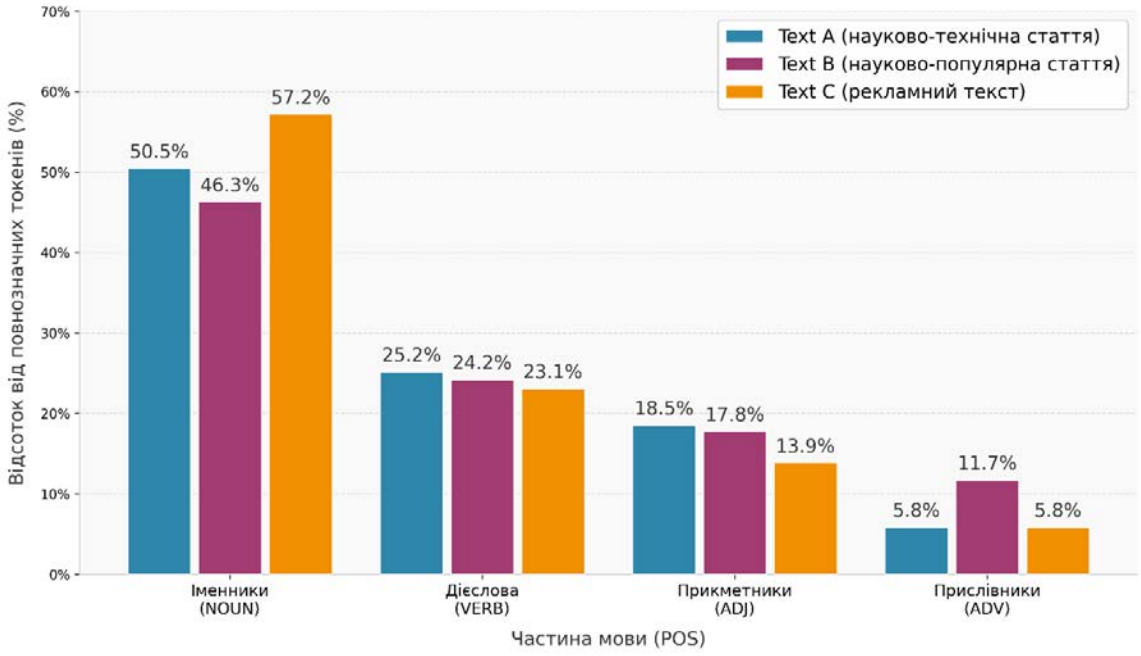

Рис. 1. Порівняння частотності частин мови в аналізованих текстах (за результатами аналізу засобами spaCy)

Науково-технічний текст є найбільш дієслівним завдяки описам методів та процедур (*we propose*, *we implement*, *we conduct*, *we demonstrate*), тоді як рекламний текст має найнижчу частку дієслів, що пояснюється його номінативним характером.

Розподіл прикметників також демонструє спадну тенденцію: A (18.51%) → B (17.77%) → C (13.86%). Відносно низька частка прикметників у тексті C є несподіваним результатом, однак пояснюваним: позитивна конотація у рекламному прес-релізі передається переважно через іменникові групи з препозитивними означеннями (*next-generation*, *award-winning*, *groundbreaking*), які spaCy класифікує як іменникові модифікатори.

Найбільш показовим жанровим маркером є розподіл прислівників: текст B різко виділяється на фоні текстів A та C (11.72% проти 5.85% і 5.84% відповідно). Висока частка прислівників є характерною рисою науково-популярного дискурсу, де вони слугують для емоційного підсилення та вираження суб'єктивного ставлення автора (*increasingly*, *significantly*, *potentially*, *dramatically*). Практично однаковий показник текстів A та C (~5.8%) свідчить про їхню схожу стилістичну нейтральність щодо прислівникового вираження.

З перекладознавчої точки зору, отримані результати дозволяють сформулювати диференційовані рекомендації щодо перекладацької стратегії. Переклад тексту A вимагає точного відтворення термінологічних конструкцій та процедурних описів при збереженні академічного стилю. Переклад тексту B передбачає особливу увагу до збереження емоційної забарвленості та позитивної конотації, що є нетиповим завданням для перекладу наукового дискурсу. Переклад тексту C потребує точної передачі іменникових груп із складними означеннями, характерних для рекламного прес-релізу.

Наступним кроком комп'ютерно-лінгвістичного дослідження стало визначення лексичної щільності – вагомого параметра, що дозволяє формалізувати та порівняти складність досліджуваних матеріалів. У дослідженнях академічного письма вона розглядається як важливий показник рівня володіння писемною мовою та стилю автора [18]. У кількісній та корпусній лінгвістиці лексична щільність є співвідношенням змістовних слів (які відрізняються від граматичних або службових слів) до загальної кількості слів у тексті. Змістовними вважаються класи слів, що несуть семантичне навантаження: такі як іменники, дієслова (за винятком модальних), прикметники та прислівники, похідні від прикметників [17].

Показники лексичної щільності для досліджуваних текстів представлені в таблиці 1.

Отримані результати демонструють неочікувану жанрову картину. Найвищу лексичну щільність має текст A (58.79%), що відповідає теоретичним очікуванням щодо науково-технічного дискурсу та узгоджується з висновками [18], де встановлено вищу лексичну щільність у точних науках порівняно з іншими дисциплінами. Однак найнижчий показник демонструє не рекламний текст C (57.14%), а науково-популярна стаття B (51.66%), що суперечить традиційним уявленням про жанрову ієрархію лексичної щільності.

Близькість показників текстів A та C (58.79% та 57.14% відповідно) пояснюється інформаційною насиченістю рекламного прес-релізу як жанру: переліки технічних характеристик та назв продуктів (*AI Skin Analysis APIs*, *Virtual Try-On APIs*, *First-Party Data Engine*, *Omnichannel Deployment*) підвищують лексичну щільність до рівня, характерного для науково-технічного дискурсу. Найнижчий показник тексту B (51.66%) зумовлений вищою питомою вагою службових слів, що забезпечують емоційну забарвленість, доступність викладу та зв'язність науково-популярного тексту, орієнтованого на широку аудиторію.

Так, лексична щільність є чутливим, але неоднозначним індикатором жанрової належності тексту: рекламний прес-реліз може демонструвати лексичну щільність, порівнянну з науково-технічним текстом, якщо він насичений термінологічними переліками та назвами продуктів. Цей результат підкреслює важливість комплексного підходу до доперекладацького аналізу, що поєднує кілька параметрів NLP-аналізу.

З точки зору перекладу, близькі показники лексичної щільності текстів A та C не означають однакової складності їхнього перекладу. Переклад тексту A вимагає максимальної термінологічної точності в передачі наукових концепцій, тоді як

Таблиця 1

Порівняння лексичної щільності в аналізованих текстах

	Text A	**Text B**	**Text C**
Загальна кількість токенів (без пунктуації та стоп-слів)	1070	1024	1288
Кількість повнозначних слів	629	529	736
Лексична щільність	58.79%	51.66%	57.14%

переклад тексту С потребує збереження рекламного ефекту та легкості сприйняття. Переклад тексту В, попри найнижчу лексичну щільність, передбачає збереження емоційної забарвленості та суб'єктивно-оцінного характеру науково-популярного дискурсу.

Окремий пласт перекладацької складності становить відтворення предметних номенклатур, що зумовлює доцільність залучення інструментів розпізнавання іменованих сутностей. Зазначена NLP-процедура полягає в автоматичній ідентифікації та класифікації категорій об'єктів у тексті.

Бібліотека spaCy є однією з найпоширеніших бібліотек для NER завдяки високій продуктивності та зручному API. Поєднання spaCy з моделлю en_core_web_trf забезпечує надійний інструмент розпізнавання іменованих сутностей у текстах різних жанрів [1]. Зазначена модель підтримує 18 категорій іменованих сутностей для англійської мови, повний перелік яких наведено в офіційній документації [14].

Аналіз іменованих сутностей засобами spaCy виявив суттєві жанрові відмінності між текстами як у кількісному, так і в якісному вимірі (таблиця 2).

Загальна кількість іменованих сутностей зростає від тексту А до тексту С (26 → 55 → 83), що відображає різну комунікативну природу жанрів. Кількість категорій NER також збільшується: текст А має лише 4 категорії, тексти В і С – 10 і 11 відповідно, що свідчить про зростання тематичного різноманіття від академічного до рекламного дискурсу.

Категорія PRODUCT домінує у тексті А (57.69%), де представлена виключно назвою досліджуваної системи Gödel Agent, що підтверджує монотематичність академічного дискурсу. У тексті С вона займає друге місце (27.71%) і охоплює назви продуктів та технологій (Skin.Chat, YouCam AI Agent, Virtual Try-On APIs). У тексті В значення PRODUCT є мінімальним (3.64%) – лише Claude Code, що характерно для науково-популярного дискурсу, який апелює до понять, а не до конкретних продуктів.

Категорія ORG демонструє протилежну тенденцію: повна її відсутність у тексті А є значущим нулем – академічний текст зосереджений на ідеях та методах, а не на організаціях. Текст В згадує організації у контексті соціально-політичного дискурсу (Anthropic, OpenAI, EU), тоді як текст С демонструє найвищу частку ORG (36.14%) завдяки частому згадуванню брендів та цифрових платформ (Perfect Corp., Haut.AI, Amazon, Facebook Messenger, Instagram, WhatsApp).

Категорія PERSON виконує різні функції в кожному з текстів. У тексті А імена математиків-теоретиків (Gödel, Schmidhuber) є носіями концептуального змісту та маркерами теоретичного підґрунтя дослідження. У тексті В категорія PERSON представлена різноманітніше: вчені (Alan Turing, Yoshua Bengio), громадські діячі (Pope Leo XIV) та дослідники (Yudkowsky), що свідчить про широкий соціальний контекст науково-популярного дискурсу. Майже повна відсутність PERSON у тексті С (1.20%) є характерною рисою рекламних текстів, де акцент робиться на продуктах та брендах.

Категорія DATE відображає часову орієнтованість текстів. Мінімальна кількість дат у тексті А (лише рік публікації теоретичного джерела) підтверджує позачасовий характер наукового дискурсу. Найвища частка DATE у тексті В (27.27%) свідчить про активне вписування подій у часовий контекст (the mid-1960s, march 2023, may 2026), що є типовою рисою науково-популярного дискурсу. Текст С містить дати переважно у контексті анонсів та комерційних заходів (CES 2026, January 6, November 20, 2025).

Показовим є також набір унікальних категорій кожного тексту. Текст В є єдиним, що містить категорії GPE (US, Ireland), NORP (British, Chinese) та PERCENT (80%, over 50%), що свідчить про його широкий соціально-географічний контекст. А текст С є єдиним, що містить категорії EVENT (CES 2026), TIME (24/7) та MONEY ($250 Amazon gift card), що є характерними маркерами реклами.

З перекладознавчої точки зору, висока частка PRODUCT та PERSON у тексті А вимагає точ-

Таблиця 2

Порівняльний аналіз іменованих сутностей аналізованих текстів

Категорія	Text A	Text B	Text C
Усього сутностей	26	55	83
Категорій NER	4	10	11
PRODUCT	15 (57.69%)	2 (3.64%)	23 (27.71%)
ORG	-	12 (21.82%)	30 (36.14%)
PERSON	7 (26.92%)	10 (18.18%)	1 (1.20%)
DATE	1 (3.85%)	15 (27.27%)	13 (15.66%)
Унікальні категорії	-	GPE, NORP, PERCENT, CARDINAL, FAC	EVENT, TIME, MONEY, FAC

ної передачі власних назв систем та імен науковців, які є носіями концептуального змісту. Найрізноманітніший набір категорій NER у тексті B (10 категорій) потребує від перекладача широкої енциклопедичної підготовки: знання історичного контексту, сучасного технологічного ландшафту та соціально-політичного дискурсу. Домінування ORG та PRODUCT у тексті C вимагає верифікації офіційних назв брендів та продуктів у цільовій мові, а наявність унікальних категорій EVENT та MONEY потребує культурної адаптації рекламних елементів.

Логічним продовженням розпізнавання іменованих сутностей є вихід на рівень синтаксичного моделювання. Аналіз структури речення є фундаментальним методом NLP. Його використовують для перетворення природної мови на структуровані та нормалізовані тексти, що дає синтаксичні ознаки для подальшої обробки [16].

У досліджуваних текстах аналіз структури речення здійснювався за трьома параметрами: середньою довжиною речення (у токенах), максимальною довжиною речення та середньою глибиною синтаксичного дерева залежностей. Перший показник дає загальне уявлення про синтаксичну складність тексту, другий – про наявність надмірно довгих конструкцій, які майже завжди потребують членування при перекладі, а третій виявляє рівень синтаксичної вкладеності. Результати подано на діаграмах (рис. 2 та рис. 3).

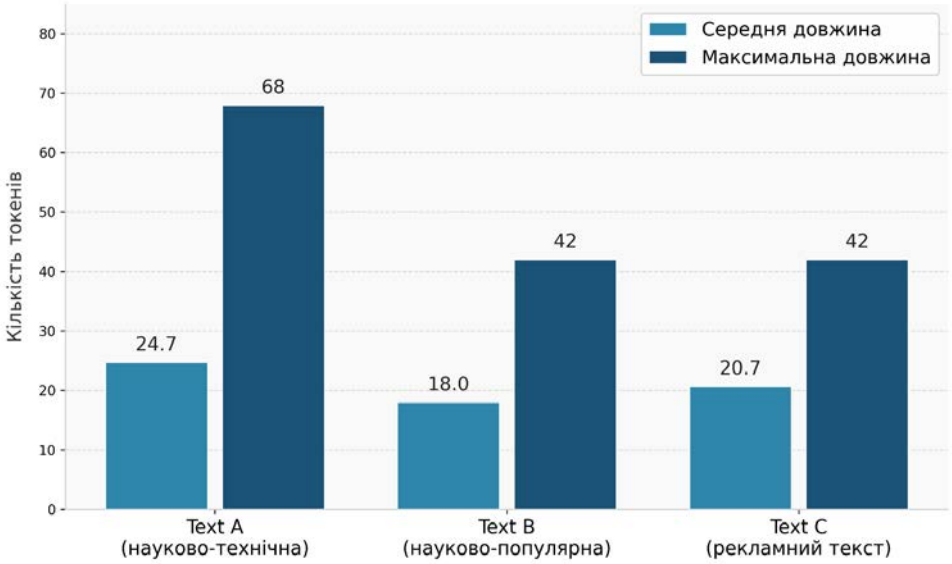

Рис. 2. Довжина речення в аналізованих текстах (токени)

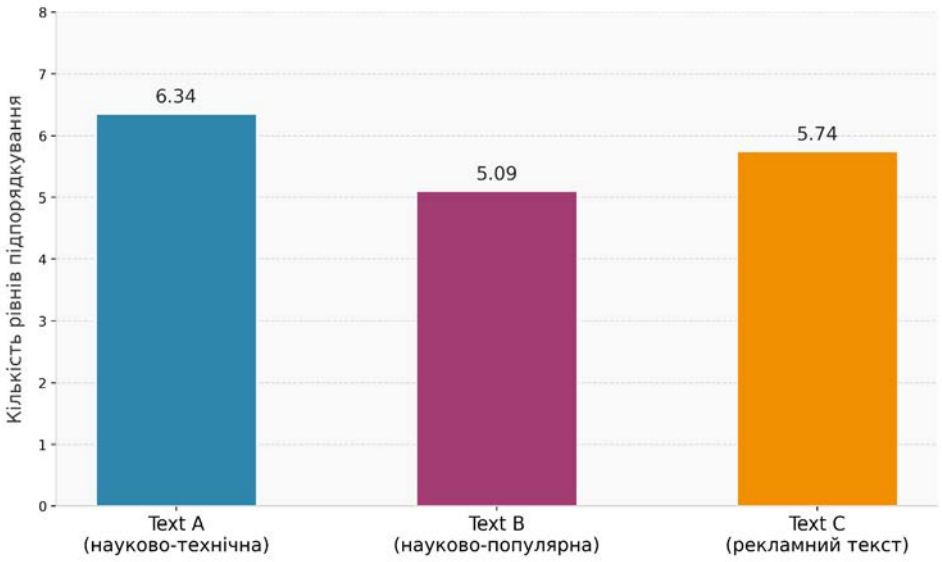

Рис. 3. Середня глибина синтаксичного дерева залежностей в аналізованих текстах

Результати демонструють чітку жанрову ієрархію за всіма трьома параметрами: текст А посідає перше місце за синтаксичною складністю, текст С – проміжну позицію, текст В має найнижчий показник, що узгоджується з висновками про кореляцію між довжиною речення та жанровою належністю тексту [6].

Текст А демонструє найвищу середню довжину речення (24.72 токени) та найглибше синтаксичне дерево (6.34). Максимальна довжина речення (68 токенів) майже вдвічі перевищує відповідний показник текстів В і С (42 токени), що свідчить про наявність складних синтаксичних конструкцій з численними рівнями підпорядкування. Текст В характеризується найкоротшими реченнями (17.98 токени) та найменшою глибиною дерева (5.09), що відповідає орієнтації науково-популярного дискурсу на широку аудиторію. Текст С займає проміжну позицію (20.67 токени та 5.74 глибина дерева). Це пояснюється чергуванням розгорнутих описових речень із короткими спонукальними конструкціями, що є характерним синтаксичним ритмом рекламного прес-релізу.

З перекладознавчої точки зору, переклад тексту А вимагає вміння відтворювати складні багаторівневі конструкції, а надмірно довгі речення (понад 50 токенів) можуть потребувати реструктуризації. Переклад тексту В передбачає збереження динамічного ритму короткого речення, тексту С – синтаксичного чергування як жанрового маркера рекламного дискурсу.

Висновки. У роботі обґрунтовано вибір Python-бібліотеки spaCy як інструменту доперекладацького аналізу та визначено три ключові параметри дослідження відповідно до моделі К. Норд. Матеріалом слугували англомовні тексти трьох жанрів на спільну тему штучного інтелекту, що дозволило розглядати жанр як єдину контрольовану змінну. Результати NLP-аналізу засвідчили, що кожен із досліджуваних жанрів має виразний лінгвістичний профіль. Виявлені жанрові відмінності визначають перекладацьку стратегію та можуть слугувати основою для формування диференційованих рекомендацій перекладачеві ще до початку власне перекладацького процесу.

Перспективи подальших досліджень вбачаються у розширенні параметрів аналізу за рахунок додаткових NLP-інструментів та збільшенні корпусу текстів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. A novel NLP-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage / S. Ferro et al. *Neural Computing and Applications*. 2025. URL: https://doi.org/10.1007/s00521-025-11449-2 (date of access: 12.08.2026).

2. Bayrak M. The Contribution of Pre-translation Reading to the Translation Process. *Uluslararası Özel Amaçlar için İngilizce Dergisi*. 2025. Vol. 3, no. 1. P. 57–63. URL: https://doi.org/10.70870/joinesp.1712788 (date of access: 12.08.2026).

3. Computer-Assisted Text Analysis for Comparative Politics / C. Lucas et al. *Political Analysis*. 2015. Vol. 23, no. 2. P. 254–277. URL: https://doi.org/10.1093/pan/mpu019 (date of access: 12.08.2026).

4. Gasparyan G. Pre-Translation Text Analysis as an Essential Step to an Effective Translation. *Armenian Folia Anglistika*. 2020. Vol. 16, no. 1 (21). P. 117–129. URL: https://doi.org/10.46991/afa/2020.16.1.117 (date of access: 12.08.2026).

5. Güzel S. Text Analysis in Translation Studies: Nord's Translation Oriented Source Text Analysis. *Internatioal theory, research and review in philology*. Ankara, 2023. P. 137–162. URL: https://surl.li/komlrz (date of access: 12.08.2026).

6. Liu K., Afzaal M. Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLOS ONE*. 2021. Vol. 16, no. 6. P. e0253454. URL: https://doi.org/10.1371/journal.pone.0253454 (date of access: 12.08.2026).

7. Macanovic A. Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*. 2022. P. 102784. URL: https://doi.org/10.1016/j.ssresearch.2022.102784 (date of access: 12.08.2026).

8. Miyata R. Formulating a terminology for source document profiling through a literature review: from functionalist to documentational approaches. *Perspectives*. 2022. P. 1–18. URL: https://doi.org/10.1080/0907676x.2022.2049830 (date of access: 12.08.2026).

9. Nord C. Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translationoriented Text Analysis. 2nd ed. Amsterdam – New York : Rodopi, 2005. 274 p.

10. Onyshchak H. V., Popovych Y. V., Pochepetska T. M. TRANSLATION TEXT ANALYSIS AS AN ESSENTIAL PART OF FUTURE TRANSLATOR'S TRAINING. *Transcarpathian Philological Studies*. 2023. Vol. 1, no. 32. P. 138–141. URL: https://doi.org/10.32782/tps2663-4880/2023.32.1.25 (date of access: 12.08.2026).

11. Poncio F. P. An investigation of the gender gap in the information technology and engineering programs through text mining. *Decision Analytics Journal*. 2022. P. 100158. URL: https://doi.org/10.1016/j.dajour.2022.100158 (date of access: 12.08.2026).

12. Seresová K., Breveníková D. The Role of Text Analysis in Translation. *CBU International Conference Proceedings*. 2019. Vol. 7. P. 617–622. DOI: 10.12955/cbup.v7.1428.

13. Short term diachronic shifts in part-of-speech frequencies / C. Mair et al. *International Journal of Corpus Linguistics*. 2002. Vol. 7, no. 2. P. 245–264. URL: https://doi.org/10.1075/ijcl.7.2.05mai (date of access: 12.08.2026).

14. spaCy. English trained pipelines: en_core_web_trf label scheme. spaCy Models Documentation. URL: https://spacy.io/models/en#en_core_web_trf-labels (date of access: 12.08.2026).

15. Spring R., Johnson M. The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools. *System*. 2022. Vol. 106. P. 102770. URL: https://doi.org/10.1016/j.system.2022.102770 (date of access: 12.08.2026).

16. Zhang X., Mao R., Cambria E. A survey on syntactic processing techniques. *Artificial Intelligence Review*. 2022. URL: https://doi.org/10.1007/s10462-022-10300-7 (date of access: 12.08.2026).

17. Zhou X., Gao Y., Lu X. Lexical complexity changes in 100 years' academic writing: Evidence from Nature Biology Letters. *Journal of English for Academic Purposes*. 2023. P. 101262. URL: https://doi.org/10.1016/j.jeap.2023.101262 (date of access: 12.08.2026).

18. Zhu H., Wang T., Pang N. Diachronic changes in lexical density of research article abstracts: A corpus-based study. *Lingua*. 2024. Vol. 312. P. 103837. URL: https://doi.org/10.1016/j.lingua.2024.103837 (date of access: 12.08.2026).

Стаття поширюється на умовах ліцензії відкритого доступу CC BY 4.0

Дата першого надходження статті до видання: 31.08.2026
Дата прийняття статті до друку після рецензування: 06.11.2026
Дата публікації (оприлюднення) статті: 24.11.2026