Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

СИНТАКСИЧНА СТИЛЕМЕТРІЯ ЖАНРІВ УКРАЇНСЬКОМОВНОЇ ДИТЯЧОЇ ЛІТЕРАТУРИ: ЕКСПЕРИМЕНТАЛЬНЕ КОРПУСНЕ ДОСЛІДЖЕННЯ

Кваліфікаційна робота
на здобуття ОС «магістр»
студентки ІІ року навчання
галузі знань 03 «Гуманітарні науки», спеціальності 035 «Філологія», спеціалізації 035.10 «Прикладна лінгвістика», ОНП «Прикладна лінгвістика (редакторсько-перекладацька та експертна діяльність)»
Вероніки Ярославівни ГІЛЬТАЙЧУК

Науковий керівник:
д.філол.н., проф.
Наталія ДАРЧУК

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від 14.05.2026 **№** 14
Завідувач кафедри
к.філол.н., доц. **Сергій РІЗНИК**

КИЇВ – 2026

АНОТАЦІЯ

Магістерська робота «Синтаксична стилеметрія жанрів українськомовної дитячої літератури: експериментальне корпусне дослідження» присвячена автоматичній синтаксичній параметризації українськомовних дитячих текстів та виявленню кількісних закономірностей їхньої жанрової організації. Актуальність зумовлена відсутністю в українській лінгвістиці систематичних стилеметричних описів синтаксичної специфіки жанрів дитячої літератури.

Об'єктом дослідження є українськомовна дитяча література, предметом – синтаксична та морфологічна структура дитячих текстів як маркер жанрової організації. Метою є автоматична граматична параметризація дитячих текстів та виявлення закономірностей їхньої жанрової організації.

Основними завданнями були: формування корпусу трьох жанрів (97 казок, 57 оповідань, 21 поетичний текст), порівняльна апробація NLP-інструментів і вибір оптимального парсера, параметризація за 15 кількісними показниками та розробка програмного забезпечення для автоматичного синтаксичного аналізу.

Методологічну основу становлять корпусна лінгвістика, кількісна стилеметрія та методи автоматичної обробки природної мови. Для синтаксичного аналізу застосовано нейронний парсер Stanza, відібраний за результатами порівняльного експерименту з UDPipe і spaCy. Інноваційність полягає у першому застосуванні комплексної синтаксичної параметризації на матеріалі дитячої літератури в українській лінгвістиці.

Результати підтвердили, що жанрова приналежність є стилеметрично значущим фактором. Поезія утворює якісно відмінний синтаксичний профіль, казки й оповідання є синтаксично близькими жанрами, єдиним розрізнювальним параметром яких є частота тире як маркер діалогічності. Розроблене програмне забезпечення може застосовуватися для автоматизованої оцінки складності текстів та укладання навчальних матеріалів.

Ключові слова: стилеметрія, синтаксичний парсинг, дитяча література, дерева залежностей, корпусна лінгвістика, жанрова диференціація, синтаксичні параметри, Stanza.

ABSTRACT

The master's paper "Syntactic Stylometry of Ukrainian- Language Children's Literature Genres: An Experimental Corpus-Based Study" focuses on the automatic syntactic parameterization of Ukrainian-language children's texts and the identification of quantitative patterns in their genre organization. The relevance of the study is determined by the absence of systematic stylometric descriptions of the syntactic specificity of children's literary genres in Ukrainian linguistics.

The object of the study is Ukrainian-language children's literature; the subject is the syntactic and morphological structure of children's texts as a marker of genre organization. The aim is the automatic grammatical parameterization of children's texts and the identification of patterns in their genre organization.

The main tasks included: building a corpus of three genres (97 fairy tales, 57 short stories, 21 poetic texts), a comparative evaluation of NLP tools and selection of the optimal parser, parameterization based on 15 quantitative features, and the development of software for automatic syntactic analysis.

The methodological framework comprises corpus linguistics, quantitative stylometry, and natural language processing methods. Syntactic analysis was performed using the Stanza neural parser, selected based on a comparative experiment with UDPipe and spaCy. The novelty of the study lies in the first application of comprehensive syntactic parameterization to children's literature in Ukrainian scholarship.

The results confirmed that genre affiliation is a stylometrically significant factor. Poetry forms a qualitatively distinct syntactic profile, while fairy tales and

short stories are syntactically similar genres distinguished only by the frequency of dashes as a marker of dialogicity. The developed software can be applied for automated readability assessment of children's texts and the compilation of educational materials.

Keywords: stylometry, syntactic parsing, children's literature, dependency trees, corpus linguistics, genre differentiation, syntactic parameters, Stanza.

ЗМІСТ

ВСТУП

Сучасна лінгвістика переживає етап інтенсивної трансформації, що пов'язаний із переходом від описових, традиційних методів дослідження мови до корпусних, автоматизованих та статистичних підходів. У центрі цього переходу опиняється стилеметрія – галузь, що вивчає стиль тексту шляхом формалізації мовних параметрів і кількісного опису мовленнєвої структури. Стилеметрія відкриває можливість розглядати текст як складну, але вимірювану систему, у якій частота лексичних одиниць, розподіл частин мови, синтаксичні конструкції та ритміко-пунктуаційні особливості формують унікальне стилістичне оформлення. Для української лінгвістики ця галузь є водночас новою та закономірно необхідною, адже поєднує між собою статистичний підхід [5; 67], комп'ютерну лінгвістику [4], частотну лексикографію [2] та практику корпусного аналізу, що лише останніми роками набуває активного розвитку [80].

Актуальність цього дослідження зумовлена передусім тим, що дитяча література – одна з найменш опрацьованих сфер у межах української стилеметрії та корпусної лінгвістики. Дитячий текст виконує важливі пізнавальні, виховні та мовленнєворозвивальні функції, формує передумови для подальшого читання, впливає на когнітивний розвиток і розвиток мовної компетентності дітей дошкільного та молодшого шкільного віку. Науковці наголошують на тому, що доступність дитячих текстів, ясність синтаксичної структури, виваженість ритму та простота мовлення є критичними для успішного формування мовного досвіду [48; 55; 58; 79]. Питанням стилістики та культури мови в широкому сенсі присвячені праці О. Пономаріва [56], В. Чабаненка [57] та І. Кочан [58]. Проте більшість досліджень дитячої літератури зосереджуються на лексичних, семантичних та комунікативних характеристиках тексту, тоді як синтаксична організація жанрів для дітей майже не піддавалася статистичному аналізу, особливо з урахуванням автоматизованих методів. Дотепер в українській лінгвістиці немає узагальнюючих праць, що комплексно описували б синтаксичні параметри таких жанрів, як народні казки,

авторські казки, оповідання чи поезія для дітей.

Паралельно з цим стрімко розвивається автоматизація лінгвістичних досліджень. Сучасні інструменти обробки природної мови дозволяють здійснювати морфологічний та синтаксичний аналіз текстів у великих обсягах, що відкриває нові можливості для стилеметрії [25]. Бібліотеки Python, такі як Stanza [21], spaCy [24] та інші інструменти, що працюють на основі Universal Dependencies [18], створюють умови для автоматичного складання синтаксичних дерев і розмітки частин мови. Попри це, дослідження, які порівнювали б ефективність зазначених інструментів саме в контексті української мови, практично відсутні. Наразі бракує експериментальних робіт, що встановлювали б точність, сильні та слабкі сторони парсерів і визначали оптимальний інструмент для синтаксичної стилеметрії українських текстів [9]. Особливо відчутним є брак таких досліджень у сфері дитячої літератури, де синтаксичні особливості безпосередньо пов'язані з рівнем доступності та віковою адаптацією тексту [16; 68].

Нарешті, ще одним чинником актуальності є стан розвитку самої української стилеметрії. Попри існування традиції статистичного вивчення мови [3; 67], більшість таких праць присвячені загальним функціональним стилям або аналізу лексики. Синтаксична стилеметрія залишається найменш дослідженою ланкою, хоча саме синтаксис вирізняється стабільністю, ритмічністю й високою інформативністю для жанрової характеристичності текстів [9; 10; 11]. При цьому українськомовні дитячі тексти не були системно досліджені з погляду синтаксичних індикаторів, таких як довжина речення, частка простих і складних конструкцій, вага частин мови (частка токенів відповідної частини мови від загальної кількості токенів у тексті) чи розподіл пунктуаційних одиниць.

Метою цього дослідження є здійснення автоматичної граматичної параметризації українськомовних дитячих текстів та виявлення морфологічних і синтаксичних закономірностей їхньої організації. Для цього ми поєднуємо методи корпусної лінгвістики, стилеметричного аналізу та автоматичного

синтаксичного парсингу.

Виконання мети передбачає реалізацію таких **завдань:**

1. опрацювання наукової літератури з теми та визначення методологічного підходу до синтаксичної стилеметрії;
2. структурування корпусу дитячих текстів за жанровими групами (казки, оповідання, поезія);
3. порівняльну апробацію сучасних інструментів автоматичного синтаксичного аналізу (Stanza, spaCy, UDPipe) та вибір оптимального парсера для подальшого дослідження;
4. формування трьох жанрових вибірок дослідження на основі корпусу kidus.db;
5. проведення морфологічного та синтаксичного аналізу кожної вибірки;
6. систематизацію отриманих даних у базі даних;
7. розрахунок статистичних індексів синтаксичної параметризації текстів;
8. здійснення жанрового порівняльного аналізу;
9. розробку програмного забезпечення для автоматичного синтаксичного аналізу та розрахунку стилеметричних параметрів текстів.

Об'єктом дослідження є українськомовна дитяча література, а **предметом** – синтаксична та морфологічна структура дитячих текстів, зокрема довжина речень, розподіл синтаксичних конструкцій, параметри дерев залежностей, частини мови та пунктуаційні індикатори як маркери жанрової організації. **Матеріалом** дослідження слугує корпус українськомовної дитячої літератури, який налічує понад дві сотні текстів різних жанрів: 98 казок (зокрема казка-притча, казка-п'єса, казкова повість), 57 оповідань, а також поетичні та фольклорні тексти – авторські вірші, колискові, щедрівки, колядки, купальські й жниварські пісні, лічилки, скоромовки та збірки мирилок (загалом 21 одиниця). Для синтаксичного стилеметричного аналізу з цього масиву виокремлено три жанрові вибірки: казки (97 текстів), оповідання (57 текстів) і поезія (21 текст).

Методологічну основу становлять статистичні підходи [5; 66; 3; 67], корпусні методи [4; 38; 39], а також сучасні методи автоматичної обробки текстів [25; 51]. Основним інструментом автоматичного морфологічного та синтаксичного аналізу слугує Stanza [21]; у порівняльному експерименті додатково використовувалися spaCy [24] та UDPipe [20].

Наукова новизна роботи полягає у першій спробі комплексної синтаксичної стилеметрії українськомовних дитячих текстів із використанням автоматизованого парсингу. **Практично** результати можуть бути застосовані у викладанні, автоматизованій оцінці складності текстів, укладанні дидактичних матеріалів та в подальших лінгвістичних експертизах.

Структура магістерської роботи складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатку.

РОЗДІЛ 1. СТИЛЕМЕТРІЯ – АКТУАЛЬНИЙ НАПРЯМ СУЧАСНОЇ ЛІНГВІСТИКИ

1.1. Стилеметрія в сучасній українській лінгвістиці: історія становлення, основні напрями та перспективи

У сучасній науці стилеметрія визначається як кількісне дослідження стилю тексту, що спирається на аналіз вимірюваних текстових характеристик, які описують стилістичні відмінності та уможливлюють ідентифікацію авторства. Е. Стамататос окреслює стилеметричний підхід як такий, що ґрунтується на текстових особливостях – маркерах індивідуального письма, – котрі піддаються статистичному й обчислювальному опрацюванню [1].

Стилеметрія як наукова дисципліна має глибоке коріння. Початок сучасного кількісного підходу до дослідження авторства заклали Ф. Мостеллер і Д. Уоллес у 1963 році, застосувавши статистичний аналіз функціональних слів для атрибуції «Федераліста» [35]. Від того часу дисципліна пройшла шлях від простих частотних підрахунків до багатовимірних моделей машинного навчання. Дж. Берроуз запропонував метод Delta, заснований на стандартизованих частотах найуживаніших слів, який став одним із найбільш відтворюваних і верифікованих підходів у сучасній стилеметрії [33]. М. Коппел, Дж. Шлер і Ш. Аргамон систематизували обчислювальні методи атрибуції авторства, виокремивши лексичні, синтаксичні та стилістичні ознаки як основні класи параметрів [34]. Д. Голмс простежив еволюцію стилеметрії від ранніх спроб кількісного опису тексту до зрілої обчислювальної дисципліни, відзначивши, що перехід від аналізу окремих текстів до корпусних досліджень докорінно змінив методологічні можливості галузі [37]. Водночас М. Едер, Я. Рибіцький і М. Кестемон розробили пакет stylo для мови R, який демократизував доступ до стилеметричних методів і зробив їх відтворюваними для широкого кола дослідників [36].

В українській лінгвістиці передумови стилеметрії формувалися ще в середині XX ст. у межах статистичного мовознавства, започатковані колективною монографією «Статистичні параметри стилів» (1967) [67],

«Частотним словником сучасної української художньої прози», укладеним колективом співробітників відділу структурно-математичної лінгвістики (укладачі: М. Муравицька, Н. Дарчук, Т. Грязнухіна, В. Критська та ін., за ред. В. Перебийніс, 1981) [2], монографією «Частотні словники та сфери їх використання» (В. С. Перебийніс, М. П. Муравицька, Н. П. Дарчук, 1985) [82], а також узагальнюючими працями В. Перебийніс зі статистичних параметрів українських текстів, зокрема дослідженням розподілу частин мови та довжини речень у різних функціональних стилях (2002) [3]. Ці роботи заклали кількісно-лінгвістичне підґрунтя, без якого розвиток власне стилеметрії був би неможливим. Паралельно розвивалася комп'ютерна лінгвістика – зокрема роботи Т. Грязнухіної, Н. Дарчук, Л. Орлової, В. Критської, а також праці Є. Карпіловської [4] з формалізованого опису морфеміки. Роботи відділу структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні визначили методологічні орієнтири для автоматизованого аналізу текстів. Від початку 2000-х рр. стилеметрія набуває самостійності як дослідницький напрям – передусім завдяки роботам В. В. Левицького [81], а також С. Бук, яка застосувала статистичні методи до порівняльного аналізу функціональних стилів української мови [5].

Сьогодні в українській стилеметрії можна виокремити три провідні напрями. Першим є кількісне дослідження функціональних стилів, спрямоване на виявлення відмінностей між науковим, масмедійним, офіційно-діловим, художнім та іншими типами текстів з погляду лексичного розподілу, різноманітності й концентрації. Порівняльний аналіз С. Бук показує, що стилістичну диференціацію можна описати статистично, розглядаючи стиль як систему повторюваних мовних виборів, що піддаються числовому опису та порівнянню між корпусами текстів [5].

Другим напрямом є стилеметрія, орієнтована на дослідження авторства, тобто вивчення індивідуального письма за допомогою кількісних характеристик. Е. Стамататос розглядає атрибуцію авторства як одну з центральних ділянок сучасних міжнародних стилеметричних досліджень. В

українському контексті цей напрям також набуває актуальності, В. Литвин, В. Висоцька, П. Пукач підтверджують, що ідентифікація авторства може ґрунтуватися на таких параметрах, як лексична різноманітність, когерентність, синтаксична складність і концентрація лексики [6]. О. В. Бісікало та О. Р. Іщенко підтверджують доцільність формалізації синтаксичних ознак речень у завданнях атрибуції із застосуванням машинного навчання: запропонований авторами набір із 11 структурних характеристик, отриманих за допомогою синтаксичного парсингу бібліотекою Stanza, забезпечив точність класифікації на рівні 92% [7].

Третім є напрям комп'ютерного й автоматизованого аналізу текстів, що відображає ширшу цифрову трансформацію в гуманітаристиці та лінгвістиці. Показовим прикладом є система TextAttributor 1.0 – інструмент автоматичного стилеметричного аналізу українських медіатекстів [8]. Система здійснює статистичну параметризацію за допомогою 15 індексів, що охоплюють лексичний, морфологічний, синтаксичний, психолінгвістичний та семантичний рівні тексту, і забезпечує як стилеметричний аналіз ідіостилю, так і атрибуцію авторства. Ця розробка підтверджує, що стилеметрія в Україні вийшла за межі теоретичних дискусій і ручних розрахунків та дедалі більше спирається на спеціалізоване програмне забезпечення й обчислювальні процедури. У цьому ж напрямі у статті «Parameterization of the Ukrainian Text Corpus based on parsing» (Н. Дарчук, В. Сорокін, 2022) [9] запропоновано систему автоматичної параметризації синтаксичної структури речень у вигляді дерев залежностей та апробовано її на матеріалі поетичного корпусу Ліни Костенко; виділено такі параметри, як кількість вузлів, глибина та ширина дерева, асиметрія та довжина дуги. У продовженні цього підходу стаття «On stylistic differentiation in Ukrainian poetic speech based on the syntactic structure of sentences» (Н. Дарчук, О. Зубань, В. Сорокін, 2024) [10] здійснює порівняльний стилеметричний аналіз ідіостилю трьох українських поетів – Ліни Костенко, Миколи Вінграновського та Івана Драча – на матеріалі 7 752 речень із застосуванням восьми параметрів графів залежностей, що мають стилерозрізнювальну силу.

Зазначені напрями засвідчують міждисциплінарний характер стилеметрії: її методи ґрунтуються на лінгвістичній стилістиці, статистиці, комп'ютерній лінгвістиці та машинному навчанні. Стилеметрію в Україні слід розглядати не як вузьку підгалузь стилістики, а як міждисциплінарний напрям, у якому лінгвістична інтерпретація поєднується з формальними, статистичними та обчислювальними процедурами.

Подальший розвиток стилеметрії в Україні пов'язаний із кількома реалістичними напрямами. По-перше, розширення цифрових текстових баз даних та анотованих корпусів відкриє можливості для ширших і надійніших кількісних досліджень – зокрема в тих сферах, де корпусна база досі залишається недостатньою, наприклад дитяча література. По-друге, вдосконалення NLP-інструментів для української мови – зокрема синтаксичних парсерів і морфологічних аналізаторів – сприятиме розвитку синтаксичної стилеметрії, яка сьогодні є найменш дослідженою ланкою порівняно з лексичним рівнем. Про перспективність цього напряму свідчить і сам колектив розробників TextAttributor, який у висновках статті окреслює плани щодо інтеграції синтаксичного та психолінгвістичного аналізу в подальші версії системи. По-третє, методологічна інтеграція лексичних, морфологічних і синтаксичних параметрів дасть змогу перейти до комплексних, багаторівневих стилеметричних описів, що матимуть як теоретичну, так і прикладну цінність, зокрема для автоматизованої оцінки складності текстів і розробки дидактичних матеріалів. Саме в цьому контексті синтаксична стилеметрія українськомовної дитячої літератури, якій присвячена ця робота, постає як актуальний і перспективний дослідницький проєкт.

Отже, стилеметрія в сучасній українській лінгвістиці виросла з кількісних підходів до мови, пройшла шлях від статистичного вивчення функціональних стилів і нині активно розвивається завдяки обчислювальним та автоматизованим методам. Три провідні напрями – кількісне порівняння стилів, дослідження авторства та автоматизований аналіз текстів – визначають її сьогоднішнє обличчя. Подальший поступ галузі залежатиме від розширення

цифрових ресурсів і вдосконалення інструментів для обробки українського тексту.

1.2. Квантитативні параметри в стилеметрії

Квантитативні параметри – основний інструмент стилеметричного аналізу, саме вони дозволяють перейти від суб'єктивного опису стилю до вимірюваних і відтворюваних характеристик. У лінгвістичному сенсі параметр – це квант інформації про мовну структуру, що разом з іншими параметрами формує числовий профіль тексту [11]. Завдяки цьому стиль розглядається не як інтуїтивне враження, а як система повторюваних мовних виборів, які можна кількісно описати і порівняти між текстами різних авторів, жанрів і стилів [1].

Морфологічні параметри. Розподіл частин мови є надзвичайно інформативним і стабільним стилеметричним індикатором. Хоча всі частини мови наявні в текстах будь-якого функціонального стилю, їх кількісне співвідношення виявляється стилістично маркованим. Кожен стиль характеризується власним частиномовним профілем, що відображає його комунікативну функцію та структурні особливості. Це було переконливо продемонстровано ще в 1967 р., коли аналіз розподілу частин мови по стилях показав, що іменник посідає перший ранг в усіх стилях, однак співвідношення дієслів, іменних форм і службових слів суттєво варіюється [3]. З часом на основі цих спостережень було розроблено систему морфологічних індексів, придатних для автоматичного обчислення: індекс іменникових означень (IIO), індекс дієслівних означень (IDO), ступінь номіналізації (STN), індекс субстантивності (ISUB) та індекс займенниковості (IPRO) [8; 11]. Ці індекси дозволяють встановити, наскільки текст є статичним чи динамічним, конкретним чи абстрактним, і виявити індивідуальні авторські вподобання на рівні граматики.

Лексичні параметри. Статистичні характеристики словника тексту утворюють окрему групу параметрів, що описують лексичне багатство,

варіативність і концентрацію. Ключовими серед них є індекс різноманітності (V/N), що відображає відношення обсягу словника до обсягу тексту, індекс винятковості тексту (V1/N) і словника (V1/V), що вимірюють частку hapax legomena, та індекс концентрації (V10т/N), який показує питому вагу високочастотної лексики [11]. Ці параметри виявляють стійкі міжстильові закономірності, наприклад поетичне мовлення характеризується найвищим індексом різноманітності, тоді як офіційно-діловий стиль – найнижчим, що відповідає інтуїтивним уявленням про ці стилі, але тепер отримує точне числове вираження [5]. Для вимірювання відстані між текстами застосовують евклідову відстань, косинусну схожість та індекс Жаккарда – метрики, що широко використовуються в стилеметричних дослідженнях [8; 11].

Синтаксичні параметри. Синтаксичний рівень є особливо значущим для стилеметрії, оскільки відображає не лише лексичний добір, а й спосіб організації думки. Довжина речення, частка простих і складних конструкцій, розподіл типів синтаксичного зв'язку – це базові параметри, що описують структурну організацію тексту і чутливо реагують на жанрові та авторські відмінності [3; 11]. Із розвитком автоматичного синтаксичного парсингу з'явилася можливість значно розширити набір синтаксичних параметрів за рахунок характеристик дерев залежностей: кількості вузлів речення, глибини та ширини дерева, асиметрії графа, довжини дуги та кількості однорідних груп [9; 10]. Ці параметри є особливо інформативними для порівняльного аналізу авторських стилів, оскільки синтаксична організація тексту відзначається відносною стабільністю і слабко піддається свідомому контролю з боку автора.

Пунктуаційні індикатори. Пунктуація утворює відносно компактну, але інформативну групу стилеметричних ознак. Частотність різних розділових знаків корелює із синтаксичними рішеннями автора, ритмікою тексту та жанровою приналежністю. Художні й розмовні тексти відрізняються вищою питомою вагою питальних і окличних конструкцій, тоді як наукові й офіційні – переважанням розповідних речень із мінімальною пунктуаційною варіативністю. У поетичному мовленні пунктуаційні ознаки нерозривно

пов'язані з ритміко-синтаксичною організацією тексту, що підтверджується порівняльним аналізом ідіостилів Л. Костенко, М. Вінграновського та І. Драча [10].

Ефективність стилеметричного аналізу суттєво зростає при поєднанні параметрів різних рівнів у єдину модель. Жоден окремо взятий параметр не є достатнім для надійної класифікації тексту – лише комплексний опис забезпечує стабільні та відтворювані результати [1]. Саме цей принцип лежить в основі сучасних стилеметричних систем, які оперують десятками індексів одночасно [8; 11].

Кількісні підходи до аналізу тексту спираються на розвинений апарат корпусної статистики. М. Окс систематизував статистичні методи, застосовні в корпусній лінгвістиці, зокрема критерії порівняння частот, міри асоціації та кластерний аналіз, що дають змогу встановлювати статистичну значущість виявлених закономірностей [52]. С. Ґріс конкретизував практичне застосування цих методів у мовознавчих дослідженнях, описавши процедури обчислення дисперсії, нормалізації та порівняння корпусів різного обсягу [53]. Питання коректного порівняння корпусів різного жанрового складу і обсягу детально розглянуто А. Кілгаріффом [54]. Застосування цих методів є обов'язковою умовою для того, щоб стилеметричні висновки не зводились до описових спостережень, а мали статистичне підґрунтя.

1.3. Корпусна лінгвістика як інструмент стилеметричних досліджень.

Розвиток корпусної лінгвістики змінив методологічний підхід до стилеметрії. Від інтуїтивного уважного читання – до емпірично обґрунтованого масштабного аналізу текстів. Зокрема, автоматизована обробка великих текстових корпусів дозволила вийти за межі дослідження окремих творів і дійти статистично обґрунтованих висновків щодо авторського та жанрового стилю. Корпус, як структурована, репрезентативна збірка текстів, закодованих у машиночитаному форматі, є дієвим аналітичним інструментом. Він дає змогу дослідникам виявляти закономірності розподілу, які залишаються непомітними

на рівні окремого тексту, та піддавати теоретичні гіпотези ретельній перевірці на статистично значущих вибірках [12]. З огляду на те, що обґрунтовані кількісні висновки щодо стилю можна зробити лише за умови систематичного аналізу достатнього обсягу текстового матеріалу, сучасна стилеметрія практично неприйнятна без методологічного підґрунтя, заснованого на корпусі.

Методологічні засади корпусної лінгвістики детально описані в класичних підручниках дисципліни. Т. МакЕнері і Е. Вілсон визначають корпус як репрезентативну машиночитану збірку текстів, зібрану відповідно до чітких критеріїв відбору, і наголошують на тому, що репрезентативність і збалансованість є ключовими вимогами до будь-якого дослідницького корпусу [38]. Д. Байбер, С. Конрад і Р. Реппен показали, що корпусний підхід дає змогу систематично описувати граматичні, лексичні та регістрові особливості мови, недосяжні при інтроспективному аналізі [39]. Д. Кеннеді в своєму огляді наголошує на тому, що корпусна лінгвістика є не просто набором технік, а принципово іншою парадигмою мовознавства – емпіричною і кількісною за своєю суттю [40]. Ще раніше Дж. Сінклер обґрунтував ідею про те, що мова функціонує через великі одиниці – колокаційні патерни та фразеологічні блоки – які можна виявити лише на матеріалі великих корпусів [41; 42; 72; 75].

Морфологічне та синтаксичне анотування корпусу дає змогу автоматично генерувати словники частотності частин мови, детальні розподіли синтаксичних конструкцій та комплексні лексичні профілі, тобто ті кількісні дані, що лежать в основі стилеметричної параметризації та надають їй емпіричного змісту. Без корпусної основи більшість стилеметричних показників залишалися б абстрактними теоретичними конструкціями, оскільки їх систематичний розрахунок для сотень текстів стає можливим лише завдяки автоматизованій обчислювальній обробці. Крім того, достатньо репрезентативний корпус дає змогу встановити статистичні норми та довірчі інтервали для кожного параметра, щодо яких можна змістовно та порівняльно оцінити ступінь відхилення, що виявляється у конкретному тексті чи автора [8].

В Україні корпусна база активно розвивалася протягом останніх десятиліть. Корпус української мови (mova.info) охоплює тексти художнього, наукового, ділового та публіцистичного стилів із морфологічними, синтаксичними та семантичними анотаціями і є основним ресурсом для стилеметричних досліджень в українській науці [15]. Саме на основі цього корпусу було систематично проведено параметризацію поетичних текстів та розроблено й удосконалено методологічні підходи до автоматичного синтаксичного аналізу [9; 10]. Паралельно з цим Генеральний регіонально анотований корпус української мови (ГРАК) [30] становить додатковий великомасштабний ресурс, що охоплює понад 600 мільйонів словоформ і надає дослідникам значні можливості для більш широких діахронічних та соціолінгвістичних досліджень варіацій [14].

Корпусна лінгвістика спирається на три ключові принципи – репрезентативність вибірки, збалансованість жанрового складу та відтворюваність аналітичних результатів [12]. Для порівняння жанрів, яке є центральним завданням нашого дослідження, репрезентативність є особливо важливою. Ідеальна кількісна збалансованість не завжди досяжна – і наш корпус це відображає: 98 казок, 57 оповідань і 21 поетичний текст. Різниця в обсягах вибірок є методологічним обмеженням і враховуватиметься при інтерпретації результатів. Анотування корпусу – присвоєння текстовим одиницям лінгвістичних тегів – є тим, що уможливлює автоматичний масштабний розрахунок морфологічних і синтаксичних параметрів [13].

Для цілей цієї роботи корпусна лінгвістика відіграє подвійну роль. Вона є водночас методологічною основою для формування дослідницького корпусу і набором обчислювальних інструментів для його обробки. Будь-який змістовний стилеметричний аналіз починається зі складання репрезентативної вибірки текстів, у нашому випадку казок, оповідань та віршів, яка і дає підстави для порівняння синтаксичних параметрів між жанрами. Накопичений в українській науці досвід корпусно-стилеметричних досліджень прози й поезії слугує для нас важливим методологічним орієнтиром [9; 10; 11].

1.4. Дитяча література як об’єкт стилеметричного аналізу

Дитяча література є багатогранною сферою лінгвістичних досліджень, що поєднує в собі художнє вираження, дидактичну мету та повсякденну комунікативну функцію. Тексти, призначені для юних читачів, стикаються з особливим завданням. Вони повинні захоплювати увагу, залишаючись при цьому цілком доступними та ретельно адаптованими до когнітивних і мовних здібностей, властивих для відповідного етапу розвитку читача. Синтаксична стисліть, чітка та передбачувана структура речень, виважена ритмічна структура та навмисне лексичне повторення – усі ці риси не є простими спрощеннями, а слугують цілеспрямованими конструктивними принципами, а не стилістичними обмеженнями. Саме вони визначають жанрову специфіку дитячої літератури.

Зв’язок між лінгвістичною структурою тексту та віком його цільового читача ґрунтується на десятиліттях психолінгвістичних досліджень. Впливова теорія когнітивного розвитку Жана Піаже передбачає [27], що діти проходять через низку окремих стадій інтелектуального розвитку, кожна з яких характеризується специфічними когнітивними здібностями, стратегіями міркування та способами сприйняття навколишнього світу. Ці стадії розвитку мають прямий вплив на мовну складність, яку тексти можуть обґрунтовано вимагати від своїх читачів: на доопераційній стадії (вік 2–7 років) діти переважно оперують конкретними, матеріальними поняттями та значною мірою покладаються на прості синтаксичні форми, тоді як на стадії конкретно-операційного мислення (вік 7–11 років) вони поступово розвивають здатність опрацьовувати складніші граматичні відношення та абстрактні ідеї. Емпіричні дані, отримані в результаті досліджень розвитку мовлення, надають додаткову підтримку цим закономірностям. Наприклад, чотирирічна дитина зазвичай володіє словниковим запасом приблизно 1500 слів і тільки починає експериментувати зі складанням складних речень, тоді як у віці 8–10 років діти регулярно використовують складні граматичні структури та демонструють здатність розуміти синтаксичні відношення значної складності [12]. Таким

чином, синтаксична організація дитячого тексту – це одночасно стилістичний і педагогічний інструмент, що безпосередньо впливає на його доступність і освітню цінність. М. Томаселло обґрунтував конструктивістський підхід до засвоєння мови дітьми на основі соціальної взаємодії [74].

Жанрова специфіка дитячої літератури, що безпосередньо визначає її синтаксичну організацію, описана в класичних фольклористичних і літературознавчих дослідженнях. В. Пропп, аналізуючи структуру чарівної казки, виявив, що вона організована за стійкою послідовністю функцій, незалежних від конкретних персонажів і сюжетних деталей [46]. Ця структурна передбачуваність є однією з причин синтаксичної простоти казкового тексту: повторюваність конструкцій, діалогічність і паратаксис є не стилістичними вадами, а жанровими константами. Л. Дунаєвська, досліджуючи українську народну казку, підкреслює її орієнтацію на усне мовлення і специфічну ритмо-синтаксичну організацію, що виражається в стійких зачинах, кінцівках і повторах [47]. А. Богуш, досліджуючи мовленнєвий розвиток дітей дошкільного і молодшого шкільного віку, підкреслює, що синтаксична простота і ритмічна повторюваність дитячих текстів є не стилістичними обмеженнями, а дидактично вмотивованими принципами, що безпосередньо впливають на засвоєння мовних норм [48]. С. Крашен у теорії мовленнєвого вводу обґрунтував, що текст є засобом розвитку мовної компетентності лише тоді, коли його складність перевищує поточний рівень читача на один крок [49; 77]. К. Хант, вивчаючи письмо дітей різного віку, встановив, що синтаксична зрілість найкраще вимірюється через середню довжину Т-одиниць – мінімальних термінальних одиниць, що відповідають одній незалежній клаузі з усіма підрядними [50]. Ця ідея є теоретичним попередником сучасних підходів до вимірювання синтаксичної складності через параметри дерев залежностей.

Кількісні методи широко та продуктивно застосовуються у дослідженні дитячої літератури, відіграючи центральну роль у розробці спеціалізованих мовних ресурсів, адаптованих для юних читачів у різних мовах та освітніх контекстах. Серед найпомітніших прикладів – Oxford Wordlist (англійська) [31],

ретельно складений на основі систематичного аналізу розмовної та писемної мови дітей віком 5–9 років; Manulex (французька) [28], лексична база даних на основі частотності, складена з підручників початкової школи для 1–5 класів; ChildLex (німецька) [29], обширний корпус, що містить 117 952 леми, взяті з 500 дитячих книг, орієнтованих на читачів віком 6–12 років; та CREA Junior (іспанська) [32], ресурс, спеціально розроблений для відображення автентичних, натуралістичних мовних моделей іспаномовних дітей [12]. Незважаючи на відмінності в мові та обсязі, ці проекти об'єднує фундаментальна опора на корпусні та кількісні методології, що чітко вписує їх у традицію стилеметричного аналізу.

На рівні власне стилеметричних досліджень дитячої літератури найбільш показовою є робота *"A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults"* у якій Ваутер Хавералс, Ліндсі Гейбельс та Ванесса Джусен дослідили авторський стиль 10 нідерландських та англомовних кросрайтерів – письменників, які пишуть одночасно для дітей і дорослих – засобами стилеметрії [13]. Результати засвідчили, що тексти для дітей демонструють стійкі відмінності у частотних профілях слів порівняно з текстами тих самих авторів для дорослих, що підтверджує – вікова адресованість тексту є стилеметрично значущим фактором, який фіксується кількісними методами. Сяо Ю., Доусон Н. Дж., Банерджі Н., Нейшн К. аналізуючи письмо дітей різного віку, встановили, що синтаксична складність є більш чутливим індикатором мовного розвитку, ніж лексична різноманітність, що додатково обґрунтовує вибір синтаксичних параметрів як центрального об'єкта цього дослідження [16].

В українській науці дитяча література досліджувалася переважно в літературознавчому та педагогічному вимірах. Лінгвістичні дослідження зосереджувалися здебільшого на лексичному рівні: зокрема, частотний словник на матеріалі українськомовної дитячої літератури засвідчив специфіку лексичного складу цього масиву і довів, що дитячий текст формує власний лексичний профіль, відмінний від художньої прози для дорослих [2]. Однак,

незважаючи на ці досягнення, синтаксична організація ключових жанрів дитячої літератури, зокрема казок, оповідань та поезії – ніколи не була предметом систематичного стилеметричного аналізу в українській науковій традиції. Це помітне й значуще упущення, особливо з огляду на те, що синтаксична доступність відіграє безпосередню та формувальну роль у мовленнєвому та когнітивному розвитку юних читачів у критичні періоди мовного розвитку. Питання читабельності тексту та доступності для різних вікових груп досліджували Р. Флеш [63], С. Кросслі [64] і П. Нейшн [65]; дослідженням частотності слів у великих корпусах присвячена праця Дж. Ліча зі співавторами [76]. Однак синтаксична організація ключових жанрів дитячої літератури, зокрема казок, оповідань та поезії, ніколи не була предметом систематичного стилеметричного аналізу в українській науковій традиції.

Попри усталену міжнародну традицію кількісного аналізу дитячих текстів і зростаючий інтерес до стилеметрії різних жанрів, українська дитяча література в синтаксичному вимірі так і не стала предметом систематичного дослідження. Це не просто прогалина в літературі – це завдання, яке цілком можливо і потрібно вирішити засобами корпусної та обчислювальної лінгвістики.

ВИСНОВКИ ДО РОЗДІЛУ 1

У цьому розділі ми заклали теоретичне підґрунтя дослідження: розглянули стилеметрію як міждисциплінарний напрям, простежили її становлення в Україні від статистичного мовознавства середини XX ст. до сучасних автоматизованих систем і виявили три провідні напрями – кількісний аналіз функціональних стилів, стилеметрію авторства та автоматизований аналіз текстів. Серед них синтаксична стилеметрія залишається найменш дослідженою.

Ми також систематизували основні групи квантитативних параметрів: морфологічні індекси, лексичні показники, синтаксичні параметри дерев залежностей та пунктуаційні індикатори. Жоден із них сам по собі не дає надійного результату – лише комплексне поєднання параметрів на основі репрезентативного анотованого корпусу забезпечує достовірні стилеметричні висновки.

Окремо ми обґрунтували роль корпусної лінгвістики як методологічної бази дослідження і розглянули основні ресурси – Корпус української мови (mova.info) і ГРАК. Дитяча ж література виявилася перспективним, але малодослідженим об'єктом стилеметрії: міжнародний досвід підтверджує, що вікова адресованість тексту кількісно вимірювана, а синтаксична складність є надійним індикатором доступності.

Таким чином, теоретичний огляд підтвердив обраний підхід – автоматичну синтаксичну параметризацію українськомовних дитячих текстів із жанровим порівнянням. Ключовим питанням, від якого залежить якість усього подальшого аналізу, є вибір інструменту синтаксичного парсингу – саме цьому присвячений наступний розділ.

РОЗДІЛ 2. СИНТАКСИЧНИЙ ПАРСИНГ УКРАЇНСЬКОМОВНОГО ТЕКСТУ: ТЕОРІЯ, МЕТОДИ, ІНСТРУМЕНТИ

2.1. Поняття синтаксичного парсингу. Історія розвитку парсерів для української мови.

Синтаксичний парсинг – це автоматичне визначення та формальне представлення граматичної структури речення. Результатом є ієрархічна структура, яка відображає синтаксичні відношення між складовими мовленнєвої одиниці. Залежно від теоретичної моделі, ця структура зазвичай представлена або у вигляді дерева безпосередніх складників, що ілюструє, як слова об'єднуються у фрази та більші синтаксичні конструкції, або у вигляді дерева залежностей, яке кодує прямі бінарні відносини між словами на основі принципу «головний-залежний» [17]. У стилеметричному аналізі дерева залежностей є особливо інформативними, оскільки вони дають змогу обчислювати параметри синтаксичної складності, такі як глибина дерева, ширина розгалуження, довжина дуги та асиметрія графа, що безпосередньо відображають структурні характеристики стилю автора [9; 10].

Розвиток сучасного синтаксичного розбору для багатьох мов, зокрема української, став можливим завдяки міжнародній системі Universal Dependencies (UD). UD надає уніфіковану схему граматичного анотування, що охоплює частини мови, морфологічні ознаки та синтаксичні залежності [18]. Її основною перевагою є міжмовна сумісність – парсери для різних мов можна навчати та оцінювати за єдиними стандартами, що полегшує порівняння результатів та перенесення моделей. Більшість сучасних інструментів для автоматичного синтаксичного аналізу було розроблено в рамках цієї системи.

Теоретичне підґрунтя синтаксичного аналізу залежностей закладено в роботах Й. Нівре, який систематизував базові поняття dependency grammar і обґрунтував переваги залежнісного підходу перед фразовим для морфологічно багатих мов із вільним порядком слів [44]. Зокрема, залежнісні структури природним чином кодують довгі залежності між словами, типові для флективних мов – таких як українська. Х. Лю показав, що середня довжина

залежності між словами є надійним психолінгвістичним індикатором складності речення для сприйняття: чим коротші дуги в дереві залежностей, тим легше текст обробляється читачем [43]. Цей зв'язок між синтаксичними параметрами дерев і когнітивною доступністю тексту є теоретичним обґрунтуванням використання параметрів залежнісного аналізу для оцінки синтаксичної складності в цьому дослідженні.

Теоретичне підґрунтя залежнісного синтаксичного аналізу закладено в класичній праці Л. Теньєра, який першим систематично описав залежнісні відношення між словами і запропонував поняття регентного вузла як організуючого центру синтаксичної структури [69]. В українській граматичній традиції синтаксичний устрій мови детально описано в монографіях А. Загнітка, де розроблено типологію синтаксичних одиниць і структур сучасної української мови [59; 78]. Паралельно в міжнародній лінгвістиці М. Геллідей розробив системно-функційний підхід до граматики, який розглядає синтаксичні структури не як самодостатні форми, а як засоби реалізації комунікативних функцій тексту [70]. І. Вихованець у своїй граматиці синтаксису також спирається на функційний принцип опису синтаксичних одиниць, що споріднює його підхід із сучасними залежнісними моделями [71].

Прогрес у синтаксичному розборі української мови суттєво сповільнювався через відсутність великих стандартизованих корпусів із синтаксичними анотаціями. Цю проблему було вирішено створенням UD Ukrainian-IU, який слугує “золотим стандартом” для синтаксичного анотування в українській мові. Розроблений Н. Коцибою, Б. Москалевським та М. Романенком у рамках проєкту Universal Dependencies [19], цей корпус охоплює широкий спектр жанрів, включаючи художню літературу, новини, журналістику, статті Вікіпедії, юридичні документи та соціальні медіа. Зараз він слугує основним ресурсом для навчання та оцінювання сучасних синтаксичних аналізаторів для української мови.

Впровадження UD Ukrainian-IU дозволило інтегрувати українську мову в багатомовні системи, сумісні з UD. UDPipe був одним із перших інструментів,

що запропонував повний цикл для автоматичної обробки тексту, включаючи токенізацію, тегування, лематизацію та синтаксичний аналіз, всі з яких були навчені на даних UD [20; 45]. За допомогою UDPipe українські тексти можна було автоматично токенізувати, морфологічно анотувати та синтаксично аналізувати в рамках єдиного, стандартизованого процесу. Згодом українська мова була включена до Stanza, нейронного циклу NLP, розробленого Stanford NLP Group, який підтримує токенізацію, лематизацію, тегування частин мови, морфологічну анотацію та синтаксичний аналіз [21]. Іншим широко використовуваним інструментом є spaCy – бібліотека NLP, що забезпечує токенізацію, морфологічний аналіз та синтаксичний аналіз за залежностями [24]. Хоча Stanza в першу чергу призначена для академічних досліджень із заздалегідь навченими нейронними моделями, spaCy робить акцент на швидкості та масштабованості. Підтримка української мови в spaCy забезпечується за допомогою моделі uk_core_news_sm, навченої на корпусі UD Ukrainian-IU, що гарантує сумісність із тим самим стандартом анотації [19; 24]. Паралельно були розроблені інструменти морфологічного аналізу, такі як pymorphy2, разом із пакетом словників pymorphy2-dicts-uk, для забезпечення морфологічного аналізу та лематизації для української мови, хоча це не є синтаксичними парсерами у прямому сенсі [22]. Ці морфологічні ресурси були необхідними передумовами для розробки комплексних синтаксичних конвеєрів.

Останнім досягненням у галузі синтаксичних ресурсів для української мови є випуск у 2025 році нового банку синтаксичних дерев на основі виступів у Верховні Раді України. Цей ресурс доповнює існуючі корпуси UD та розширює жанрове охоплення даних із синтаксичною анотацією [23]. Цей крок знаменує перехід від опори на єдиний базовий корпус до створення більш різноманітної екосистеми ресурсів із синтаксичною анотацією, що охоплюють різні жанри та стилі.

Отже, розвиток синтаксичних аналізаторів для української мови пройшов шлях від морфологічних словників через стандартизоване анотування UD до сучасних нейронних систем – і цей шлях іще не завершено.

2.2. Методи аналізу: rule-based, dependency parsing, constituency parsing.

Синтаксичний аналіз тексту здійснюється за допомогою кількох методів, кожен з яких характеризується певними теоретичними засадами, форматами вихідних даних та рівнями автоматизації. У стилеметричних дослідженнях вибір методу синтаксичного аналізу має ключове значення, оскільки він впливає як на точність синтаксичного розбору, так і на різноманітність параметрів, доступних для кількісної оцінки. Сучасна комп'ютерна лінгвістика виділяє три основні підходи: аналіз на основі правил *(з англ. rule-based)*, синтаксичний аналіз за складовими *(з англ. constituency parsing)* та синтаксичний аналіз за залежностями *(з англ. dependency parsing)*. Кожен підхід розвивався незалежно та має свої переваги й обмеження [25; 26].

Підхід на основі правил, найдавніший метод автоматичного синтаксичного аналізу, спирається на формальні граматичні правила, які вручну кодують лінгвісти. Ці системи аналізують речення, послідовно застосовуючи правила, що регулюють поєднання частин мови, порядок слів та синтаксичні конструкції. Головною перевагою цього підходу є його прозорість та передбачуваність, кожне рішення щодо розбору можна відстежити до конкретного правила, що полегшує лінгвістичну верифікацію. Однак системи на основі правил мають значні обмеження, зокрема проблеми зі структурною неоднозначністю, залежність від повноти граматики та значні вимоги до обслуговування. Ці проблеми особливо гострі для мов з багатою морфологією, таких як українська, де вільний порядок слів та множинність граматичних інтерпретацій для однієї словоформи роблять системи на основі правил особливо вразливими [17; 25].

Обмеження підходів на основі правил спонукали до розробки статистичних методів синтаксичного аналізу, які використовують ймовірнісні моделі, навчені на великих анотованих корпусах, замість явних правил. Перехід від синтаксичного аналізу на основі правил до статистичного наприкінці 1980-х та на початку 1990-х років став значною зміною парадигми в обробці природної

мови, оскільки мову почали моделювати як статистичну систему, де певні конструкції зустрічаються з різною частотою [25]. Статистичні синтаксичні аналізатори, зокрема ті, що використовують методи умовних випадкових полів (CRF) та опорних векторів (SVM), враховують контекстну інформацію для вибору найімовірнішого синтаксичного тлумачення серед альтернатив. У залежному синтаксичному аналізі статистичні методи реалізуються у двох основних архітектурах: на основі переходів (інкрементне побудовування дерева через послідовність переходів) та на основі графів (глобальна оцінка дерева з використанням алгоритмів максимального покриття) [26]. Пізніше впровадження архітектур нейронних мереж, включаючи рекурентні мережі (LSTM) та моделі трансформерів, ще більше покращило точність розбору. Сучасні нейронні аналізатори, такі як Stanza та spaCy, навчаються на стандартизованих корпусах Universal Dependencies (UD) і досягають рівнів точності, що перевищують показники систем на основі правил, зберігаючи при цьому лінгвістичну інтерпретованість [21; 24; 25]. М. Ковінгтон запропонував один із базових детерміністичних алгоритмів залежнісного парсингу [73].

Структурна багатозначність – одне речення допускає кілька однаково правильних синтаксичних тлумачень – залишається серйозним викликом для всіх синтаксичних аналізаторів. Ця проблема особливо гостро стоїть у українській мові через вільний порядок слів та складну морфологію; одна і та ж словоформа може виступати іменником, прикметником, дієсловом або дієприкметником, що ускладнює автоматичне визначення синтаксичних відношень. Системи на основі правил вирішують проблему неоднозначності шляхом ієрархічного застосування правил, статистичні системи покладаються на ймовірності конструкцій у навчальному корпусі, а нейронні системи використовують контекстуальні представлення слів, щоб врахувати ширший контекст речення. Проте жоден із цих підходів не забезпечує повної точності [17; 25; 26].

Розбір на складники моделює синтаксичну структуру як дерево безпосередніх складників, де вузли представляють синтаксичні категорії, такі як

іменникова фраза (NP), дієслівна фраза (VP) та речення (S). Цей метод добре описує фразову організацію речення і є теоретично обґрунтованим у межах генеративної граматики. Він дозволяє ідентифікувати структурну ієрархію в складних реченнях і демонструє, як менші синтаксичні одиниці поєднуються, утворюючи більші. Однак для стилеметричних застосувань цей метод становить практичні виклики, адже вони ускладнюють автоматичний розрахунок кількісних параметрів і менш придатні для морфологічно багатих мов із вільним порядком слів. Порівняльні дослідження показують, що для флективних мов залежно-синтаксичні структури точніше відображають граматичні відношення, ніж дерева фраз [25; 26].

Метод залежностей є найпоширенішим методом у сучасній комп'ютерній лінгвістиці та стилеметрії, що представляє структуру речення у вигляді дерева залежностей. У цьому підході вузли відповідають окремим словам, а спрямовані дуги вказують на конкретні синтаксичні відношення, такі як підмет (nsubj), прямий додаток (obj), атрибут (amod), адвербіальний модифікатор (advmod) та відмінок. Кожне речення організовано як з'єднане дерево з одним кореневим вузлом, до якого всі інші слова пов'язані прямо чи опосередковано. Головною перевагою дерев залежностей для стилеметрії є їхнє пряме відображення граматичних відношень, що полегшує автоматичний розрахунок кількісних параметрів [26]. Стилеметричні параметри в цьому дослідженні – включаючи глибину дерева, ширину розгалуження, довжину дуги, асиметрію графа, кількість рівнів та частку кінцевих вузлів – всі походять з дерев залежностей [9; 10]. Крім того, синтаксичний аналіз залежностей ефективніше обробляє довгі залежності та складні синтаксичні конструкції, типові для літературних текстів, оскільки моделі на основі графів оцінюють все дерево як ціле, а не покроково [26].

Комп'ютерна лінгвістика використовує стандартизовані показники для оцінки якості синтаксичного аналізатора. Основний показник, Unlabeled Attachment Score (UAS), вимірює частку лем, для яких аналізатор правильно ідентифікує головне слово, незалежно від типу відношення. Більш суворий

показник, Labeled Attachment Score (LAS), вимагає як правильного головного слова, так і правильної мітки синтаксичного відношення, щоб лема вважалася правильно проаналізованим. LAS є основною метрикою для порівняння синтаксичних аналізаторів у рамках Universal Dependencies, оскільки вона оцінює як структурну, так і лінгвістично інтерпретовану точність [18; 25; 26].

Сучасні системи синтаксичного аналізу спираються на статистичні методи обробки природної мови, теоретичні основи яких описані в роботі К. Меннінга і Г. Шютце – одному з найвпливовіших підручників у галузі обчислювальної лінгвістики [51]. Автори детально розглядають перехід від методів правил до статистичних методів і обґрунтовують, чому імовірнісні моделі, навчені на анотованих корпусах, перевершують ручні граматики на реальних текстах. Практичним утіленням цих ідей стали спільні завдання CoNLL 2017 і 2018, де різні системи синтаксичного аналізу порівнювались на матеріалі понад 60 мов у форматі Universal Dependencies [60; 61]. Результати цих змагань показали, що нейронні системи, зокрема Stanza, стабільно перевершують статистичні попередники, що й обґрунтовує вибір нейронних парсерів для стилеметричного аналізу в цьому дослідженні.

Синтаксичний підкорпус Корпусу української мови на порталі mova.info відображає залежний підхід для української мови. У цьому корпусі структура речень анотується за допомогою іменованих типів відношень між парами слів, включаючи дієслівні та іменникові фрази (як прийменникові, так і неприйменникові), координаційні та підрядні відношеня, однорідні фрази, дієприкметникові та прислівникові фрази, а також фрази з частками [9; 11; 13]. Ця система анотації дотримується основного принципу опису синтаксичних відношень між парами слів за допомогою іменованих дуг залежності.

Наочний приклад цього підходу надає синтаксичне дерево для речення «Отже , не потрібно витрачати енергію на його відведення.», побудоване за допомогою порталу mova.info (рис. 2.1). Кольорові стрілки між словами вказують на типи синтаксичних відношень: VR (вербально-безприйменникова фраза), PP (прийменникова фраза), P-P (фраза з часткою) та NP

(іменниково-безприйменникова фраза). Водночас система автоматично обчислює кількісні параметри дерева, зокрема кількість вузлів (8), кількість рівнів графа (5), ширину розгалуження кореня (3), асиметрію графа (2/5) та середню глибину гілки (3). Ці параметри безпосередньо відповідають стилеметричним індексам, описаним Н. Дарчук та В. Сорокіним [9; 10], і є центральними для синтаксичної параметризації в цьому дослідженні.

Рис. 2.1. Синтаксичне дерево залежностей речення «Отже, не потрібно витрачати енергію на його відведення» та його кількісні параметри

2.3. Практичне застосування NLP-інструментів для синтаксичного аналізу українськомовних текстів

Вибір конкретного інструменту для синтаксичного аналізу визначає не лише точність отриманих дерев залежностей, а й формат вихідних даних, зручність інтеграції в дослідницький пайплайн та можливість автоматичного обчислення стилеметричних параметрів. У цьому підрозділі ми розглядаємо чотири основні інструменти, доступні для автоматичного опрацювання українськомовних текстів – UDPipe, Stanza, spaCy і pymorphy2 – з погляду їхніх практичних можливостей та обмежень для завдань синтаксичної стилеметрії.

Формат CoNLL-U як стандарт вихідних даних. Усі три повні NLP-пайплайни – UDPipe, Stanza і spaCy – виводять результати синтаксичного аналізу у форматі CoNLL-U, що є стандартом Universal Dependencies [18]. Цей формат представляє кожне речення як послідовність рядків, де кожен рядок відповідає одному токену і містить десять полів: індекс токена, словоформу, лему, UPOS (універсальна частина мови), XPOS (мовоспецифічна частина мови), морфологічні ознаки, індекс головного слова (HEAD), тип синтаксичного відношення (DEPREL), а також поля для додаткових залежностей і міток. Саме поля HEAD і DEPREL є основою для побудови дерева залежностей і обчислення стилеметричних параметрів [25; 26]. Стандартизація формату дозволяє використовувати єдиний код для обчислення параметрів незалежно від того, яким парсером було проаналізовано текст – це суттєво спрощує порівняльне дослідження різних інструментів.

UDPipe. UDPipe є першим повним UD-сумісним пайплайном із підтримкою української мови і доступний через Python API, командний рядок і вебінтерфейс [20]. Моделі UDPipe для української мови натреновані на UD Ukrainian-IU і реалізують токенізацію, лематизацію, POS-тегування і dependency parsing у межах єдиного конвеєра. Перевагою UDPipe є легкість розгортання і мінімальні вимоги до обчислювальних ресурсів – модель завантажується один раз і обробляє тексти без підключення до мережі. Обмеженням є те, що UDPipe використовує статистичні моделі без глибоких нейронних мереж, що знижує його точність порівняно з нейронними парсерами, особливо на нестандартних конструкціях і жанрово специфічних текстах [25]. Для художньої прози та фольклорних текстів дитячої літератури, де часто зустрічаються незвичні синтаксичні конструкції, інверсії та діалогові вставки, ця різниця може бути відчутною.

Stanza. Stanza – нейронний NLP-пайплайн Stanford NLP Group – є на сьогодні одним із найточніших інструментів для синтаксичного аналізу великої кількості мов [21]. Для кожного компонента пайплайну використовуються окремі нейронні моделі, навчені на UD-корпусах, що забезпечує високу точність

на тестових наборах UD Ukrainian-IU. Stanza надає зручний Python API: результатом обробки тексту є структурований об'єкт, де для кожного токена доступні лема, UPOS, морфологічні ознаки, індекс головного слова і тип синтаксичного відношення – саме ті поля, що необхідні для обчислення параметрів дерева залежностей. Вивід Stanza може бути легко конвертований у формат CoNLL-U для подальшої обробки. Головним обмеженням Stanza є відносно повільна швидкість обробки через нейронну архітектуру, що може бути суттєвим при роботі з великими корпусами [21; 25].

spaCy. spaCy – промислова NLP-бібліотека з акцентом на швидкість і масштабованість [24]. Підтримка української мови реалізована через модель uk_core_news_sm, натреновану на UD Ukrainian-IU, що забезпечує сумісність із тим самим стандартом анотації [19]. spaCy обробляє текст значно швидше за Stanza і надає зручний об'єктно-орієнтований API для ітерації по реченнях і токенах. Для конвертації виводу spaCy у формат CoNLL-U використовується додатковий пакет spacy-conll, що дозволяє інтегрувати її в стандартний пайплайн. Порівняно зі Stanza, спеціалізована модель spaCy для української показує дещо нижчу точність синтаксичного аналізу, особливо на морфологічно складних конструкціях і нестандартних жанрових текстах [25].

pymorphy2. pymorphy2 є словниковим морфологічним аналізатором і генератором, що використовує великий електронний словник на основі даних LanguageTool [22]. На відміну від трьох попередніх інструментів, pymorphy2 не виконує синтаксичного аналізу і не будує дерев залежностей – він зосереджений виключно на морфологічному рівні: визначенні частини мови, граматичних категорій і початкової форми слова без урахування контексту. Підтримка української мови позначена розробниками як експериментальна, що означає обмежене покриття певних словоформ і власних назв [22]. Водночас pymorphy2 відрізняється надзвичайно високою швидкістю обробки і мінімальними вимогами до ресурсів, що робить його корисним допоміжним інструментом для швидкого морфологічного маркування великих корпусів. У контексті цього

дослідження pymorphy2 може використовуватися на етапі попередньої обробки для порівняння з результатами повних пайплайнів.

Порівняльна характеристика інструментів. Чотири описані інструменти займають різні ніші в екосистемі NLP для української мови. Їхні ключові характеристики з погляду завдань цього дослідження систематизовано в таблиці 2.1.

Табл. 2.1. Порівняльна характеристика NLP-інструментів для синтаксичного аналізу українськомовних текстів

Характеристика	**UDPipe**	**Stanza**	**spaCy**	**pymorphy2**
Синтаксичний парсинг	Так	Так	Так	Ні
Морфологічний аналіз	Так	Так	Так	Так
Лематизація	Так	Так	Так	Так
Формат виводу	CoNLL-U	CoNLL-U / об'єкт	об'єкт / CoNLL-U*	об'єкт
Архітектура	Статистична	Нейронна	Нейронна	Словникова
Модель для укр. мови	UD Ukrainian-IU	UD Ukrainian-IU	UD Ukrainian-IU	LanguageTool
Точність	Середня	Висока	Середня-висока	–

Швидкість	Висока	Низька	Дуже висока	Дуже висока
Підтримка укр. мови	Повна	Повна	Повна	Експериментальна

**CoNLL-U – стандартний формат представлення результатів синтаксичного аналізу, у якому кожне слово отримує морфологічні теги та інформацію про синтаксичні зв'язки. spaCy не підтримує цей формат нативно – для його отримання використовується додаткова бібліотека spacy-conll.*

Як видно з таблиці 2.1, три інструменти – UDPipe, Stanza і spaCy – забезпечують повний синтаксичний пайплайн для української мови на основі корпусу з синтаксичною розміткою UD Ukrainian-IU, тоді як pymorphy2 обмежується морфологічним аналізом і не виконує парсингу. Ключова відмінність між трьома парсерами полягає в архітектурі: UDPipe використовує статистичну модель і є найшвидшим, Stanza – нейронну з найвищою точністю, spaCy – нейронну з найвищою швидкістю серед нейронних моделей. Для завдань стилеметричного аналізу, де точність парсингу є пріоритетом, ці відмінності мають вирішальне значення.

Розвиток інфраструктури Universal Dependencies не зупинився на першій версії анотаційного стандарту. Й. Нівре та колектив розробників у 2020 році описали другу версію UD як постійно зростаючу багатомовну колекцію корпусів із синтаксичною розміткою, яка охоплює понад 90 мов і забезпечує єдиний стандарт для навчання та оцінювання парсерів [62]. Для цього дослідження принциповим є те, що обидва основні інструменти – Stanza і spaCy – навчені на одному і тому самому корпусі UD Ukrainian-IU, що забезпечує коректність порівняння їхніх результатів і унеможливлює вплив відмінностей у навчальних даних на підсумки експериментального порівняння в підрозділі 2.4.

Для завдань синтаксичної стилеметрії, де потрібна точна побудова дерев залежностей для сотень текстів, найбільш релевантними є Stanza і spaCy як

інструменти, що надають повні дерева залежностей у UD-форматі. Вибір між ними є питанням балансу між точністю і швидкістю – і саме це є предметом експериментального порівняння в підрозділі 2.4.

2.4. Експериментальне порівняння точності парсерів

Вибір інструменту синтаксичного аналізу є одним із ключових методологічних рішень у будь-якому стилеметричному дослідженні, оскільки саме парсер визначає якість і надійність вхідних даних для подальшого обчислення параметрів. Розходження між парсерами у визначенні меж речень, синтаксичних відношень чи частин мови може призвести до систематичних похибок, які накопичуються в процесі масового аналізу корпусу й суттєво спотворюють кінцеві результати. Перед основним аналізом корпусу дитячої літератури ми провели порівняльний експеримент, метою якого є встановлення того, наскільки узгоджено три обрані парсери (Stanza, spaCy та UDPipe) обробляють українськомовні тексти різних жанрів, і який із них демонструє найвищу стабільність та надійність результатів.

Для порівняння ми обрали три тексти з корпусу ***kidus.db***, що представляють три основні жанри дослідження: казку «Сила» (Перська казка, 2440 символів), вірш «Грицеві курчата» Олександра Олеся (4831 символ) та оповідання «Свято мого дитинства» Віктора Близнеця (8680 символів). Жанрова різноманітність вибірки має вирішальне значення, оскільки різні жанри характеризуються різною синтаксичною організацією: казковий текст насичений діалогами й тире, поетичний – короткими рядками й своєрідною пунктуацією, оповідання – складними синтаксичними конструкціями. Саме на цих відмінностях найчіткіше проявляються слабкі місця кожного парсера.

Для кожного тексту і кожного парсера ми обчислювали повний набір із 14 параметрів, що охоплюють морфологічний рівень (середня довжина речення, ваги частин мови), структуру речень (частки простих/складних, двоскладних/односкладних, кількість клауз) та синтаксичні параметри дерев залежностей (глибина, ширина гілкування, довжина дуги, асиметрія, частка термінальних вузлів). Окремо фіксувався розподіл розділових знаків. Аналіз

здійснювався автоматично засобами Python із використанням бібліотек stanza, spacy та spacy-udpipe.

Морфологічні параметри відображають частиномовний склад тексту. Структурні параметри речень показують співвідношення простих і складних, двоскладних і односкладних конструкцій та середню кількість клауз – предикативних одиниць із власним присудком. Параметри дерева залежностей характеризують синтаксичну складність: глибина – кількість рівнів підрядності, ширина гілкування – середню кількість залежних від одного вузла, довжина дуги – відстань між словом і його головним словом у реченні, асиметрія – нерівномірність розподілу залежностей, частка термінальних вузлів – частку слів без залежних.

Казка. Найпомітнішою відмінністю є кількість виділених речень: Stanza і UDPipe визначили 33 речення, тоді як spaCy – лише 27 (табл. 2.2). Ручна перевірка результатів сегментації показала, що spaCy некоректно визначає межі речень у конструкціях з прямою мовою – він об'єднує ввідне речення, питальну репліку і репліку-відповідь в одну одиницю, наприклад: «Звернувся він до льоду і питає: – О льоде, звідки в тебе стільки сили? – Хіба це сила!» spaCy розглядає як єдине речення. Ця відмінність у стратегії сегментації є принциповою, оскільки безпосередньо впливає на всі похідні параметри: середня довжина речення у spaCy виявляється суттєво завищеною (17,07 проти 12,70), так само завищується кількість клауз (2,74 проти 2,15) і частка складних речень (85,2% проти 72,7%). Відповідно, синтаксичні параметри дерев у spaCy також є систематично вищими – ширина гілкування (5,67 проти 4,88), довжина дуги (3,38 проти 2,99) – що є прямим наслідком системної помилки сегментації, зумовленої жанровою специфікою навчального корпусу моделі uk_core_news_sm. Ця модель навчена переважно на новинних текстах, де діалоги з тире практично відсутні, тому тире як маркер межі речення в художній прозі модель не розпізнає коректно Стосовно морфологічних параметрів, усі три парсери демонструють близькі значення: вага іменників (частка іменників

від загальної кількості токенів у тексті) коливається в межах 0,261–0,290, вага дієслів(частка дієслів від загальної кількості токенів у тексті) – 0,169–0,195.

Табл. 2.2. Порівняльні результати синтаксичного аналізу казки «Сила»

Параметр	**Stanza**	**spaCy**	**UDPipe**
МОРФОЛОГІЧНІ ПАРАМЕТРИ			
Середня довжина речення (токени)	12,70	17,07	12,70
Вага іменників	0,290	0,261	0,263
Вага дієслів	0,195	0,169	0,189
Вага прикметників	0,011	0,017	0,021
Вага сполучників	0,060	0,065	0,073
Вага прийменників	0,118	0,124	0,126
СТРУКТУРА РЕЧЕНЬ			
Всього речень	33	27	33
Прості речення	27,3%	14,8%	33,3%
Складні речення	72,7%	85,2%	66,7%
Двоскладні речення	87,9%	92,6%	87,9%
Односкладні речення	12,1%	7,4%	12,1%
Середня кількість клауз	2,15	2,74	2,15
СИНТАКСИЧНІ ПАРАМЕТРИ ДЕРЕВА			
Глибина дерева (рівні)	3,58	3,67	3,73
Ширина гілкування в кореня	4,88	5,67	4,82

Середня довжина дуги	2,99	3,38	2,96
Асиметрія графа	0,681	0,670	0,642
Частка термінальних вузлів	0,662	0,629	0,682

Отже, на матеріалі казки ключова розбіжність між парсерами стосується сегментації речень. spaCy виділяє на 6 речень менше через некоректну обробку прямої мови – діалогічні репліки, відокремлені тире, він схильний об'єднувати з наступним реченням, що тягне за собою завищення середньої довжини речення (17,07 проти 12,70 у Stanza) і похідних синтаксичних параметрів. Морфологічні показники у всіх трьох парсерів є близькими, що підтверджує узгодженість їхніх морфологічних моделей.

Поезія. На матеріалі поетичного тексту картина змінюється (табл. 2.3). Тут уже Stanza і spaCy демонструють близьку кількість речень (127 і 128), тоді як UDPipe суттєво занижує показник до 114. Поетичний текст характеризується короткими рядками, нестандартною пунктуацією і частими переносами, що ускладнює сегментацію. UDPipe, будучи статистичною моделлю, гірше справляється з такими конструкціями – він об'єднує суміжні короткі рядки в одне речення, що призводить до завищення середньої довжини речення (7,15 проти 6,40 у Stanza) і похідних синтаксичних параметрів. Натомість Stanza і spaCy на поетичному тексті дають узгоджені результати по більшості параметрів, що підтверджує вищу надійність нейронних моделей на нестандартних жанрових текстах. Зазначимо, що поетичний текст загалом характеризується суттєво меншою середньою довжиною речення (6,40 проти 12,70 у казці), вищою часткою простих речень (67,7% проти 27,3%) і меншою глибиною дерева (2,31 проти 3,58) – що відповідає лінгвістичним очікуванням щодо синтаксичної організації поезії.

Табл. 2.3. Порівняльні результати синтаксичного аналізу вірша «Грицеві курчата»

Параметр	Stanza	spaCy	UDPipe
МОРФОЛОГІЧНІ ПАРАМЕТРИ			
Середня довжина речення (токени)	6,40	8,11	7,15
Вага іменників	0,262	0,200	0,248
Вага дієслів	0,205	0,160	0,202
Вага прикметників	0,054	0,043	0,059
Вага сполучників	0,082	0,057	0,081
Вага прийменників	0,094	0,074	0,095
СТРУКТУРА РЕЧЕНЬ			
Всього речень	127	128	114
Прості речення	67,7%	63,3%	61,4%
Складні речення	32,3%	36,7%	38,6%
Двоскладні речення	55,9%	53,1%	58,8%
Односкладні речення	44,1%	46,9%	41,2%
Середня кількість клауз	1,51	1,55	1,61
СИНТАКСИЧНІ ПАРАМЕТРИ ДЕРЕВА			
Глибина дерева (рівні)	2,31	2,69	2,62
Ширина гілкування в кореня	3,71	3,76	3,78
Середня довжина дуги	2,20	2,05	2,25
Асиметрія графа	0,547	0,500	0,543

Частка термінальних вузлів	0,646	0,564	0,645

Отже, на поетичному тексті проблемним є UDPipe, який виділяє на 13 речень менше за Stanza – статистична модель не розпізнає короткі поетичні рядки як окремі речення і об'єднує їх, звідси завищена середня довжина речення (7,15 проти 6,40 у Stanza). Нижча глибина дерева (2,31) і вища частка простих речень (67,7%) порівняно з казкою підтверджують синтаксичну простоту поетичного тексту – коротке речення з мінімальною підрядністю є відносною нормою для поезії.

Оповідання. На матеріалі оповідання всі три парсери демонструють найвищу взаємну узгодженість (табл. 2.4): кількість речень становить 131, 129 і 133 відповідно – розходження в межах 2%. Морфологічні параметри також є надзвичайно близькими: вага іменників коливається в діапазоні 0,204–0,229, вага дієслів – 0,181–0,200, вага прийменників – 0,072–0,073. Синтаксичні параметри дерева також виявляють стабільність: глибина 2,95–3,04, довжина дуги 2,58–2,65. Це вказує на те, що жанр оповідання з його стандартною синтаксичною організацією і помірною насиченістю діалогами є найменш проблематичним для автоматичного парсингу.

Табл. 2.4. Порівняльні результати синтаксичного аналізу оповідання «Свято мого дитинства»

Параметр	**Stanza**	**spaCy**	**UDPipe**
МОРФОЛОГІЧНІ ПАРАМЕТРИ			
Середня довжина речення (токени)	10,82	11,57	10,66
Вага іменників	0,225	0,229	0,204
Вага дієслів	0,198	0,181	0,200

Вага прикметників	0,094	0,093	0,104
Вага сполучників	0,081	0,079	0,083
Вага прийменників	0,073	0,072	0,072
СТРУКТУРА РЕЧЕНЬ			
Всього речень	131	129	133
Прості речення	44,3%	42,6%	39,8%
Складні речення	55,7%	57,4%	60,2%
Двоскладні речення	45,0%	49,6%	48,1%
Односкладні речення	55,0%	50,4%	51,9%
Середня кількість клауз	2,56	2,53	2,65
СИНТАКСИЧНІ ПАРАМЕТРИ ДЕРЕВА			
Глибина дерева (рівні)	3,04	2,95	3,00
Ширина гілкування в кореня	4,11	4,43	4,12
Середня довжина дуги	2,58	2,65	2,58
Асиметрія графа	0,653	0,626	0,630
Частка термінальних вузлів	0,645	0,627	0,649

Порівняльний аналіз стабільності парсерів. Узагальнюючи результати по трьох жанрах, можна виявити чіткі закономірності щодо сильних і слабких сторін кожного парсера (табл. 2.5). Таблиця добре показує, що кожен парсер має свою проблемну зону: spaCy занижує кількість речень у казці через об'єднання діалогових конструкцій, UDPipe – у вірші через труднощі з поетичною

пунктуацією. Stanza у всіх трьох жанрах дає результати, що або збігаються з іншим парсером, або є проміжними між ними.

Табл. 2.5. Кількість виділених речень по жанрах і парсерах

Жанр	Stanza	spaCy	UDPipe	Розходження
Казка	33	27	33	spaCy −18%
Вірш	127	128	114	UDPipe −10%
Оповідання	131	129	133	всі близькі

Вибір парсера для основного дослідження. За результатами порівняльного експерименту для автоматичного синтаксичного аналізу корпусу дитячої літератури ми обрали Stanza. Ця перевага обумовлена кількома факторами. По-перше, Stanza показує коректну сегментацію речень по всіх трьох жанрах, тоді як spaCy має системну помилку на діалогічних текстах з прямою мовою, а UDPipe – на поетичних текстах з короткими рядками. По-друге, нейронна архітектура Stanza на основі глибоких рекурентних мереж забезпечує вищу точність dependency parsing для морфологічно багатої мови, якою є українська, порівняно зі статистичними моделями UDPipe [21; 25]. По-третє, Stanza є найбільш широко верифікованим інструментом для завдань синтаксичного аналізу українських текстів у науковій літературі – зокрема, саме Stanza використовувалась у дослідженні О. В. Бісікало та О. Р. Іщенко для атрибуції авторства українськомовних текстів і забезпечила точність класифікації на рівні 92% [7]. UDPipe і spaCy ми використовували як контрольні інструменти – для перевірки результатів сегментації і підтвердження надійності обраного підходу.

ВИСНОВКИ ДО РОЗДІЛУ 2

У цьому розділі ми розглянули теоретичні та практичні аспекти синтаксичного парсингу стосовно завдань стилеметричного аналізу українськомовних текстів.

Ми окреслили три основні підходи до синтаксичного аналізу – на основі правил, constituency parsing і dependency parsing – і показали, що саме підхід на основі залежностей є найбільш придатним для стилеметрії: він надає структуровані дерева залежностей, кількісні параметри яких безпосередньо піддаються статистичному аналізу. Ключовою передумовою для парсингу української мови стало створення корпусів із синтаксичною розміткою, UD Ukrainian-IU, який слугує «золотим стандартом» анотації і на якому навчені всі три розглянуті нейронні та статистичні моделі [19].

Ми описали чотири NLP-інструменти – UDPipe, Stanza, spaCy і pymorphy2 – і систематизували їхні характеристики з погляду завдань дослідження. pymorphy2 виключено з порівняння як такий, що не виконує синтаксичного парсингу; UDPipe, Stanza і spaCy утворюють повні пайплайни з dependency parsing у форматі CoNLL-U.

Порівняльний експеримент на матеріалі трьох текстів різних жанрів виявив жанрово зумовлені слабкі місця кожного інструменту: spaCy некоректно сегментує діалоги з прямою мовою, UDPipe об'єднує короткі поетичні рядки. Stanza виявилася єдиним інструментом із стабільною сегментацією в усіх трьох жанрах, що і визначило її вибір як основного інструменту подальшого аналізу [7; 21].

РОЗДІЛ 3. АВТОМАТИЧНА ПАРАМЕТРИЗАЦІЯ УКРАЇНСЬКОМОВНИХ ТЕКСТІВ ДИТЯЧОЇ ЛІТЕРАТУРИ

3.1. Характеристика корпусу українськомовної дитячої літератури

Матеріальною основою дослідження слугує корпус українськомовної дитячої літератури kidus.db, сформований у межах попередньої кваліфікаційної роботи автора та розширений для потреб магістерського дослідження [12]. База даних налічує 198 текстів різних жанрів і авторів – як українських, так і зарубіжних, включаючи казки народів світу різного етнічного походження. Для цілей синтаксичного стилеметричного аналізу з цього масиву виокремлено три жанрові вибірки: казки (97 текстів), оповідання (57 текстів) та поезія (21 текст) – загалом 175 одиниць.

Ми відбирали тексти за трьома основними критеріями: відповідність цільовій віковій групі (переважно до 7 років включно), загальнодоступність джерела та рекомендованість у нормативних і методичних документах. Значну частину корпусу складають тексти зі списків рекомендованої літератури Міністерства освіти і науки України для дошкільних закладів [12], зокрема з освітніх програм «Я у Світі» та «Впевнений Старт», а також з рекомендаційного списку порталу Barabooka, укладеного Центром досліджень літератури для дітей та юнацтва. Тексти завантажувалися двома методами: автоматизованим веб-скрапінгом засобами Python (бібліотеки requests та BeautifulSoup) та ручним вилученням з відкритих джерел. Усі тексти зберігалися у форматі .txt з уніфікованою номенклатурою назв файлів (автор – назва – рік – жанр), що забезпечило автоматичну екстракцію метаданих при завантаженні до бази даних.

Вибірка казок є найчисельнішою і налічує 97 текстів 48 авторів (492 272 слововживання, 45 990 речень). Жанровий склад вибірки є різноманітним: переважна більшість текстів належить до канонічної казки (90 одиниць), а решту становлять казка-оповідання, казка-притча, казка-п'єса та казка-гра. За походженням вибірка поєднує народні та авторські твори. Народні казки представлені текстами різного етнічного походження: американські народні

казки (20 текстів), казки народів Півночі (7 текстів), англійські (2 тексти), а також перські, шотландські, німецькі, філіппінські, індійські, китайські, бразильські та арабські – по одному тексту. Серед авторських казок найповніше представлені твори Анни Казаліс (7 текстів), Шарля Перро (4 тексти), Олени Цегельської, Романа Завадовича, Емми Андієвської та Юлії Хандожинської (по 3 тексти кожного). Середній обсяг тексту у вибірці становить 474 речення, хоча розкид є значним: від 4 до 9530 речень, що відображає різницю між короткими народними казками і розлогими авторськими казковими повістями.

Вибірка оповідань налічує 57 текстів 18 авторів (209 147 слововживань, 18 506 речень). На відміну від казок, жанровий склад є однорідним – усі тексти атрибутовані як оповідання. Авторський склад охоплює як класиків зарубіжної літератури (Артур Конан Дойль – 12 текстів, Джек Лондон – 9 текстів, Редьярд Кіплінг – 4 тексти), так і українських письменників (Віктор Близнець – 6 текстів, Роман Завадович – 5 текстів, Катерина Перелісна та Всеволод Нестайко – по 4 тексти, Іван Франко та Леонід Полтава – по 2 тексти). Середній обсяг тексту становить 325 речень при діапазоні від 24 до 1134 речень, що вказує на більшу однорідність вибірки порівняно з казками.

Вибірка поезії є найменш чисельною і налічує 21 текст 16 авторів (40 361 слововживання, 5 791 речення). Жанровий склад є найрізноманітнішим серед трьох вибірок: вірші та збірки віршів (9 текстів), вірші-мирилки (2 тексти), колискові (2 тексти), колядки та щедрівки (2 тексти), а також ігрова поезія, лічилки, скоромовки, жниварські та купальські пісні – по одному тексту. Серед авторів представлені Олександр Олесь, Ліна Костенко, Андрій Малишко, Олена Пчілка, Іван Андрусяк, Грицько Бойко та інші. Значну частину вибірки (6 текстів) складають твори невстановленого або колективного авторства. Середній обсяг тексту становить 276 речень.

Загальна кількість речень у трьох вибірках – 70 287, з яких 45 990 належать казкам, 18 506 – оповіданням і 5 791 – поезії. Для казок та оповідань такий обсяг є цілком достатнім для статистично обґрунтованих висновків. Щодо поезії – 21 текст є нижнім порогом: в корпусній лінгвістиці оптимальним

мінімумом вважається 30 і більше одиниць [52], тому висновки по поетичній вибірці матимуть попередній характер і потребуватимуть верифікації. Повісті, змішані збірки та фольклорні пісенні тексти до аналізу не включалися, оскільки їхня жанрова природа виходить за межі нашого тристороннього порівняння.

3.2. Методика виділення синтаксичних параметрів. Автоматичний синтаксичний аналіз українськомовних дитячих текстів

Автоматичний синтаксичний аналіз корпусу ми здійснювали за допомогою бібліотеки Stanza – нейронного парсера Стенфордської лабораторії NLP, навченого на корпусі із синтаксичною розміткою, UD Ukrainian-IU [21]. Вибір Stanza обґрунтовано в підрозділі 2.4: серед трьох порівнюваних інструментів (Stanza, spaCy та UDPipe) саме вона показала найстабільнішу сегментацію речень в усіх трьох жанрах – правильно опрацьовувала діалоги з тире в казках і прозі та поетичні рядки, з якими spaCy схильна об'єднувати кілька рядків в одне речення, а UDPipe – занижувати загальну їх кількість.

Аналіз ми виконували скриптом stanza_analysis.py мовою Python, розробленим спеціально для цього дослідження. Скрипт послідовно зчитував тексти з бази даних kidus.db, передавав кожен текст чотириетапному конвеєру Stanza і зберігав результати у структурованому вигляді. Перший етап – токенізація – розбивав текст на окремі токени з урахуванням особливостей української орфографії та пунктуації. Другий етап – сегментація речень – визначав межі речень на основі нейронної моделі, навченої на анотованих текстах. Третій етап – морфологічний аналіз – присвоював кожному токену частиномовний тег у форматі Universal POS та лему. Четвертий етап – синтаксичний розбір – будував дерево залежностей кожного речення, визначаючи для кожного токена головне слово (head) і тип синтаксичного зв'язку (deprel) відповідно до системи Universal Dependencies [18].

Для кожного тексту ми обчислювали повний набір параметрів на основі побудованих дерев залежностей. Класифікація речень за структурою здійснювалась автоматично на основі інформації про синтаксичні зв'язки:

речення вважалося простим, якщо воно не містило підрядних клаузів, і складним – якщо містило хоча б один підрядний або сурядний клауз, виражений відповідним типом залежності у системі UD. Двоскладні речення визначалися як такі, що мають і підмет (nsubj), і присудок (root), тоді як односкладні – речення з відсутнім підметом або безособові конструкції. Середня кількість клаузів на речення обчислювалась як відношення загальної кількості клаузів до кількості речень у тексті.

Параметри дерев залежностей обчислювалися для кожного речення окремо, після чого усереднювалися по тексту. Глибина дерева визначалась як максимальна довжина шляху від кореневого вузла (root) до найвіддаленішого листового вузла. Ширина гілкування – як середня кількість безпосередніх залежних на один вузол. Середня довжина дуги – як середня лінійна відстань між головним словом і його залежним у словах. Асиметрія графа обчислювалась як відношення кількості лівих залежностей до загальної кількості залежностей. Частка термінальних вузлів визначалась як відношення кількості вузлів без залежних (листів) до загальної кількості вузлів у дереві.

Результати зберігалися у таблицях stanza_results (дані по кожному тексту окремо) та stanza_summary (зведені середні значення по жанрових вибірках) бази даних kidus.db, а також експортувалися у форматах CSV та JSON для подальшої візуалізації та статистичного аналізу.

Перед основним аналізом усього корпусу проводилася перевірка аномалій (текстів із нетиповими значеннями параметрів, що різко відрізняються від решти вибірки) методом трьох стандартних відхилень (3σ): для кожного параметра і кожної вибірки виявлялися тексти, значення яких відхиляються від середнього більше ніж на три стандартні відхилення. Виявлені аномалії не вилучалися з вибірки, однак бралися до уваги при інтерпретації результатів – зокрема тексти з екстремально великою кількістю речень можуть помітно впливати на середні значення параметрів. Для зниження впливу таких викидів основним показником порівняння слугували середні значення параметрів,

обчислені по кожному тексту окремо і потім усереднені по вибірці, а не агреговані підрахунки по всьому масиві.

Морфологічна розмітка Stanza надала інформацію про частиномовну приналежність кожного токена відповідно до системи Universal POS-тегів стандарту UD [18]. На її основі для кожного тексту обчислювалася відносна частка іменників (NOUN), дієслів (VERB), прикметників (ADJ), сполучників (CCONJ та SCONJ) і прийменників (ADP) від загальної кількості токенів. Вибір саме цих п'яти частиномовних груп зумовлений їхньою стилерозрізнювальною силою, підтвердженою в попередніх стилеметричних дослідженнях українськомовних текстів [9; 10; 11]: іменники й прикметники є маркерами номінального стилю, дієслова – динамічності тексту, сполучники – синтаксичної складності, прийменники – аналітичності морфологічної структури.

Окремо від морфологічного та синтаксичного аналізу здійснювався підрахунок пунктуаційних знаків – ком, тире та крапок – нормалізований до 1000 слів тексту. Ці показники обчислювалися безпосередньо з тексту без залучення парсера, оскільки пунктуаційні знаки в деяких випадках не зберігаються як окремі токени в дереві залежностей або інтерпретуються по-різному залежно від типу тексту. Нормалізація до 1000 слів забезпечила порівнянність текстів різного обсягу в межах і між жанровими вибірками. Двокрапки не включалися до пунктуаційного аналізу через їхню неоднозначну функцію в текстах різних жанрів.

Отримані дані пройшли візуалізацію засобами бібліотек Matplotlib та Seaborn: для кожного параметра побудовано ящикові діаграми (boxplots) по трьох жанрових вибірках, що відображають медіану, міжквартильний розмах та викиди. Зведена теплова карта (heatmap) нормалізованих значень усіх параметрів дає змогу порівняти синтаксичні профілі жанрів одночасно за всіма вимірами. Стовпчасті діаграми відображають структуру речень за типами (прості/складні, двоскладні/односкладні). Ці візуалізації є основою для порівняльного аналізу в підрозділах 3.3 і 3.4.

Програмний код дослідження розміщено у відкритому доступі *(Додаток 1).*

3.3. Автоматична статистична параметризація синтаксичної структури текстів

Автоматичну параметризацію синтаксичної структури текстів ми здійснювали за п'ятнадцятьма кількісними параметрами, згрупованими у чотири блоки: морфологічні параметри (6 показників), структура речень (1 показник; додатково – розподіл речень за типами), параметри дерев залежностей (5 показників) та пунктуаційні індикатори (3 показники). Для кожного параметра обчислювалося середнє значення (M) і стандартне відхилення (σ) по кожній жанровій вибірці. Результати аналізу наведено у таблицях 3.1–3.4 та проілюстровано відповідними рисунками.

Морфологічні параметри описують частиномовний склад текстів та середню довжину речення. Зведені значення по трьох жанрових вибірках наведено у Таблиці 3.1.

Таблиця 3.1. Морфологічні параметри жанрових вибірок (M ± σ)

Параметр	**Казки**	**Оповідання**	**Поезія**
Середня довжина речення (слів)	10,58 ± 3,69	11,15 ± 2,06	8,22 ± 2,63
Вага іменників (%)	22,99 ± 4,46	22,32 ± 3,22	30,38 ± 6,07
Вага дієслів (%)	20,10 ± 2,98	18,63 ± 2,12	18,10 ± 3,80
Вага прикметників (%)	6,22 ± 2,64	6,72 ± 1,68	6,34 ± 4,10
Вага сполучників (%)	8,77 ± 1,89	9,24 ± 1,17	5,55 ± 2,20
Вага прийменників (%)	9,15 ± 1,98	9,89 ± 1,29	9,46 ± 3,89

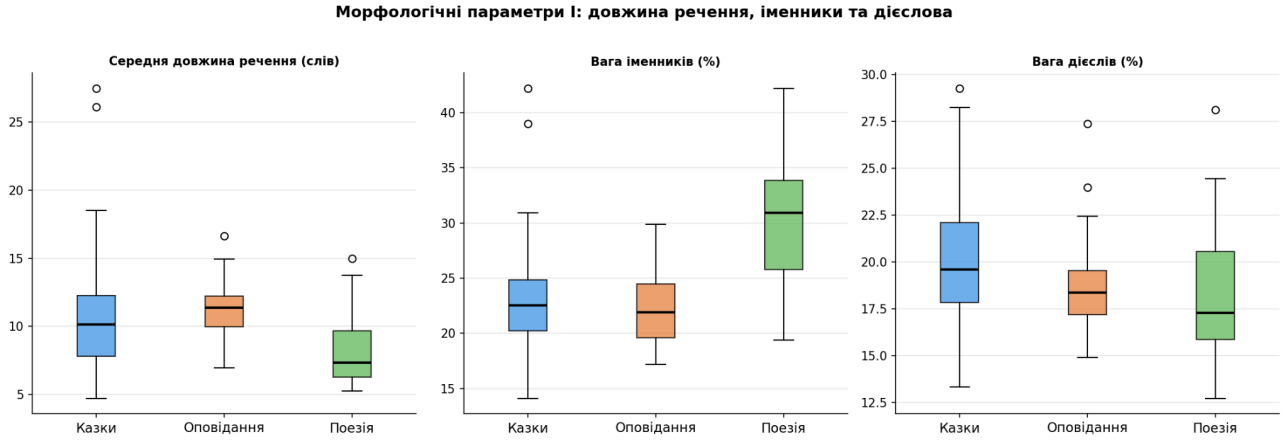

Рис. 3.1. Морфологічні параметри I: довжина речення, іменники та дієслова

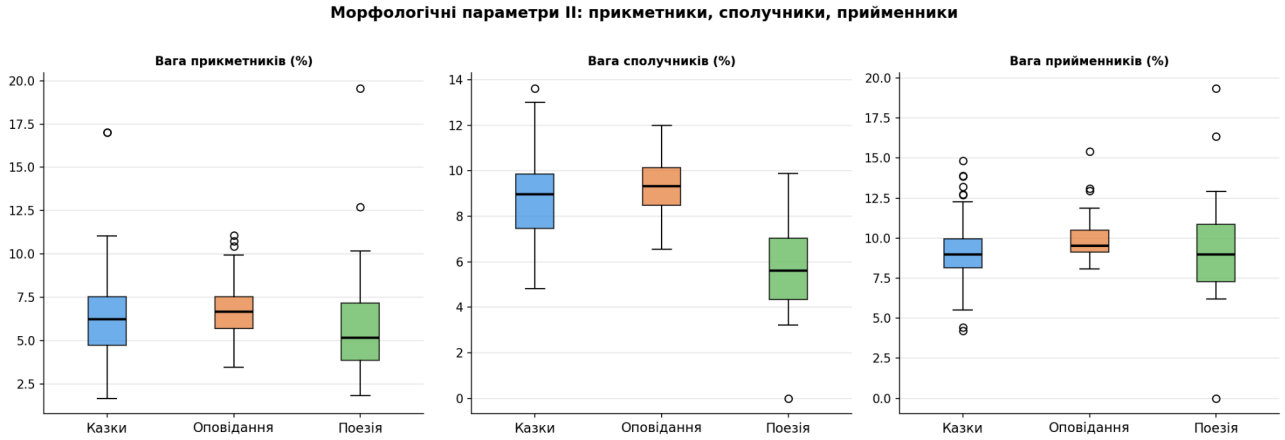

Рис. 3.2. Морфологічні параметри II: прикметники, сполучники, прийменники

Середня довжина речення є найбільш інформативним і широко застосовуваним параметром у стилеметричних дослідженнях. К. Хант обґрунтував, що середня довжина термінальних одиниць є надійним індикатором синтаксичної зрілості тексту [50], а В. Перебийніс показала, що цей параметр стабільно диференціює функціональні стилі української мови [3; 67]. У вибірці казок середня довжина речення становить 10,58 слова (σ=3,69), в оповіданнях – 11,15 слова (σ=2,06), тоді як у поезії – лише 8,22 слова (σ=2,63). Поетичні тексти мають речення коротші на 22,4% порівняно з казками і на 26,3% порівняно з оповіданнями, що пояснюється строфічною організацією вірша: кожен рядок або пара рядків, як правило, становлять синтаксично

завершену або напівзавершену одиницю. Казки й оповідання є близькими за цим параметром – різниця між ними становить лише 5,4%. Показово, що стандартне відхилення в казках (3,69) є майже вдвічі більшим, ніж в оповіданнях (2,06), що підтверджує вищу внутрішньожанрову варіативність казкових текстів – від коротких народних казок-анекдотів до розгорнутих авторських казкових повістей. Ящикові діаграми (рис. 3.1) наочно відображають цю різницю: медіана казок і оповідань є близькою, тоді як медіана поезії помітно нижча, а розкид у казках – найширший.

Вага іменників відображає номінальну щільність тексту. Поезія демонструє різко вищу вагу іменників – 30,38% (σ=6,07) – порівняно з казками (22,99%, σ=4,46) і оповіданнями (22,32%, σ=3,22). Відмінність між поезією і казками становить +32,2%, між поезією і оповіданнями – +36,1%. Натомість казки й оповідання є практично нерозрізненними: різниця між ними лише 2,9%. Висока іменникова щільність поезії пояснюється природою дитячого поетичного тексту: він будується переважно на образах-предметах, активно використовує називні речення, перелічення і звертання, де іменник є центральним елементом. Стандартне відхилення у поезії (6,07) є найвищим серед трьох жанрів і майже вдвічі перевищує відповідний показник в оповіданнях (3,22), що відображає значну варіативність між різними поетичними піджанрами – від колискових до лічилок.

Вага дієслів відображає динамічність і подієвість тексту. Казки демонструють найвищу частку дієслів – 20,10% (σ=2,98), що пояснюється їхньою подієвою динамікою: казковий наратив будується на послідовності дій і вчинків персонажів. В оповіданнях цей показник нижчий на 7,3% – 18,63% (σ=2,12), а в поезії – найнижчий: 18,10% (σ=3,80), що на 9,9% менше, ніж у казках. Відмінність між поезією і оповіданнями є незначною (-2,8%). Вища частка дієслів у казках підтверджує їхній наративний, дієвий характер; нижча в поезії – тяжіння до статичних, описових конструкцій.

Вага прикметників є єдиним морфологічним параметром, за яким усі три жанри демонструють близькі значення: 6,22% у казках (σ=2,64), 6,72% в

оповіданнях (σ=1,68) і 6,34% у поезії (σ=4,10). Жодна з міжжанрових відмінностей не перевищує 7,9% – це найменша міжжанрова варіативність серед усіх морфологічних параметрів. Ця відносна рівномірність показує, що атрибутивна функція є спільною рисою всіх жанрів дитячої літератури незалежно від їхньої структурної організації. Водночас стандартне відхилення у поезії (4,10) є у 2,5 рази більшим, ніж в оповіданнях (1,68), що відображає значну варіативність у межах поетичної вибірки – від текстів з мінімальною кількістю означень до насичено-описових.

Вага сполучників залишається одним із найбільш жанродиференціюючих морфологічних параметрів. Сполучники є безпосереднім маркером синтаксичної складності: їхня висока частка вказує на переважання складних речень із підрядним і сурядним зв'язком [9; 11]. Поезія відображає різко нижчу частку сполучників – 5,55% (σ=2,20) – порівняно з казками (8,77%, σ=1,89) і оповіданнями (9,24%, σ=1,17). Відмінність між поезією і казками становить -36,7%, між поезією і оповіданнями – -39,9% – це найбільша відносна міжжанрова різниця серед усіх морфологічних параметрів. Низька частка сполучників у поезії пояснюється переважанням простих речень і паратаксичних конструкцій, де синтаксичний зв'язок організовується ритмічно, а не через граматичні сполучники. Казки й оповідання є близькими за цим параметром – різниця між ними становить лише 5,4% на користь оповідань.

Вага прийменників відображає аналітичність синтаксичного зв'язку. Значення є відносно стабільними в усіх трьох жанрах: 9,15% у казках (σ=1,98), 9,89% в оповіданнях (σ=1,29) і 9,46% у поезії (σ=3,89). Жодна з міжжанрових відмінностей не перевищує 8,1%, що вказує на відсутність вираженої жанрової специфіки за цим параметром. Зауважимо також, що стандартне відхилення у поезії (3,89) є майже втричі більшим, ніж в оповіданнях (1,29), що знову підтверджує внутрішню неоднорідність поетичної вибірки.

Структура речень описується середньою кількістю клаузів як основним числовим параметром; додатково аналізується розподіл речень за типами:

частки простих і складних, двоскладних і односкладних. Зведені значення наведено у Таблиці 3.2.

Таблиця 3.2. Структура речень у жанрових вибірках

Параметр	**Казки**	**Оповідання**	**Поезія**
Прості речення (%)	34,2	34,5	54,8
Складні речення (%)	65,8	65,5	45,2
Двоскладні речення (%)	79,5	78,6	50,0
Односкладні речення (%)	20,5	21,4	50,0
Середня к-сть клауз (M ± σ)	2,46± 0,66	2,44 ± 0,33	2,10 ± 0,84

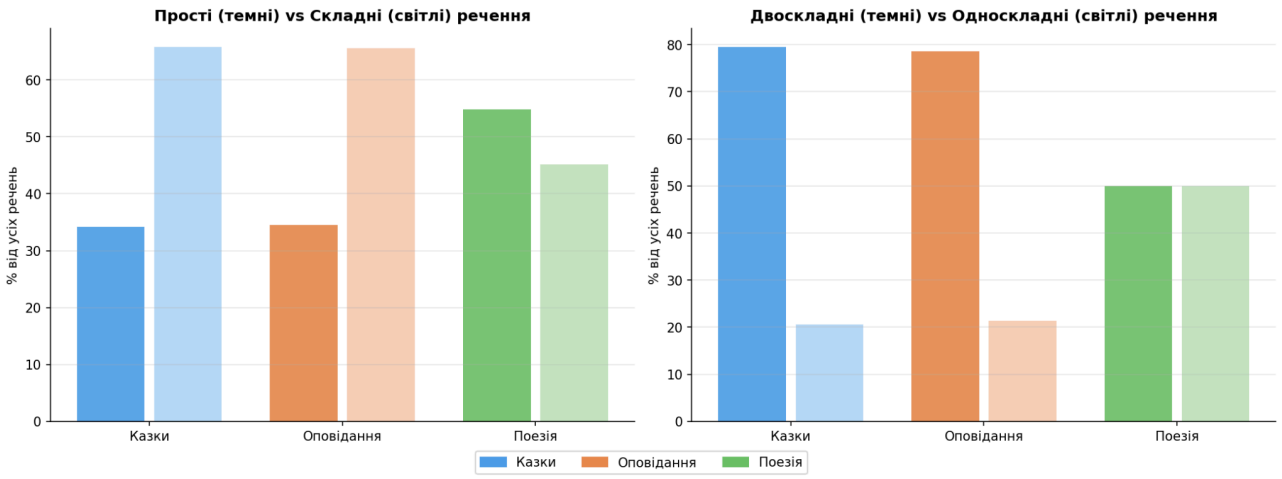

Рис. 3.3. Структура речень: прості/складні та двоскладні/односкладні

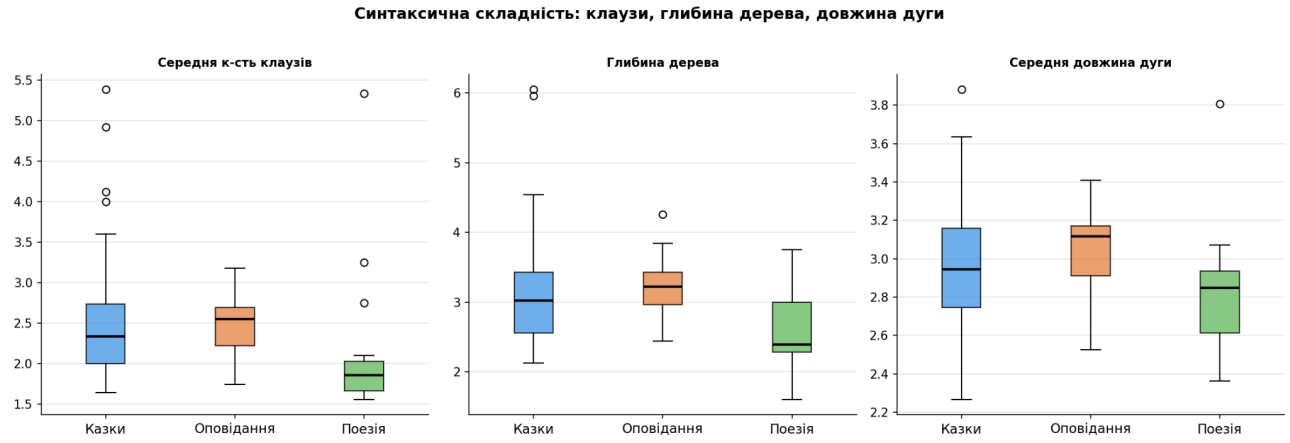

Рис. 3.4. Синтаксична складність: клаузи, глибина дерева, довжина дуги

Розподіл простих і складних речень є найбільш показовим параметром дослідження. У вибірках казок і оповідань співвідношення є практично ідентичним: близько третини речень є простими (34,2% і 34,5% відповідно) і дві третини – складними (65,8% і 65,5%). Натомість у поезії картина принципово інша: 54,8% простих речень і лише 45,2% складних. Поезія має на 60,2% більше простих речень відносно своїх складних, тоді як у прозі складні речення переважають удвічі. Ця відмінність є найбільш наочною на стовпчастій діаграмі (рис. 3.3) і підтверджує, що синтаксична організація поетичних текстів докорінно відрізняється від прозових. Переважання складних речень у казках і оповіданнях пояснюється потребою в розгорнутому оповідному синтаксисі – причиново-наслідкових зв'язках, умовних конструкціях, відносних підрядних реченнях. Поезія, навпаки, тяжіє до синтаксичної стислості й паратаксису.

Розподіл двоскладних і односкладних речень виявляє ще більш разючу відмінність між поезією і прозою. У казках двоскладні речення становлять 79,5%, односкладні – 20,5%; в оповіданнях – 78,6% і 21,4% відповідно. Тобто в обох прозових жанрах приблизно чотири з п'яти речень мають повну підмет-присудкову структуру. У поезії ж спостерігається рівне співвідношення – 50 на 50: половина речень є двоскладними, половина – односкладними. Висока частка односкладних речень у поезії пояснюється активним використанням номінативних, безособових та інфінітивних конструкцій, характерних для ліричного мовлення і дитячих піджанрів – колискових, лічилок, скоромовок.

Такі конструкції є типовим засобом поетичної стислості й економії мовних засобів. Ця відмінність є лінгвістично значущою і корелює з теоретичними уявленнями про жанрову специфіку поетичного синтаксису.

Середня кількість клаузів на речення (середня кількість клаузів на речення відображає ступінь синтаксичної складності тексту, що більше клауз, то розгалуженіша підрядна структура) підтверджує загальну тенденцію: казки (2,46 ± 0,66) і оповідання (2,44 ± 0,33) є практично нерозрізненними за цим параметром – різниця між ними становить лише 0,5%, тоді як поезія показує нижче значення (2,10 ± 0,84), що на 14,5% менше порівняно з казками і на 14,0% – з оповіданнями. Стандартне відхилення в казках (0,66) удвічі перевищує відповідний показник оповідань (0,33), що знову відображає вищу внутрішньожанрову варіативність. Слід відзначити, що стандартне відхилення у поезії є найвищим (0,84), що підтверджує значні відмінності між поетичними піджанрами за кількістю клаузів.

Параметри дерев залежностей описують геометричні та структурні характеристики синтаксичних дерев і є ключовими параметрами синтаксичної стилеметрії відповідно до методологічного підходу, розробленого Н. Дарчук та В. Сорокіним [9; 10]. Зведені значення наведено у Таблиці 3.3.

Таблиця 3.3. Параметри дерев залежностей (M ± σ)

Параметр	**Казки**	**Оповідання**	**Поезія**
Глибина дерева	3,11 ± 0,68	3,21 ± 0,39	2,62 ± 0,56
Ширина гілкування	4,28 ± 0,45	4,29 ± 0,26	3,62 ± 0,54
Середня довжина дуги	2,95 ± 0,29	3,06 ± 0,20	2,83 ± 0,30
Асиметрія графа	0,44 ± 0,04	0,44 ± 0,02	0,51 ± 0,09
Частка термінальних вузлів	0,66 ± 0,02	0,66 ± 0,01	0,64 ± 0,04

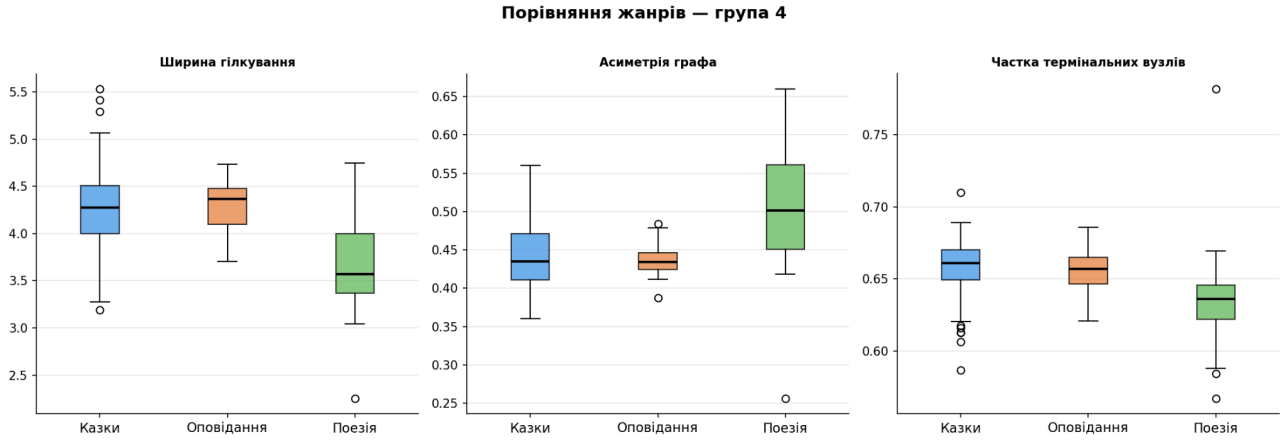

Рис. 3.5. Структура синтаксичного дерева: гілкування, асиметрія, термінальні вузли

Глибина дерева залежностей відображає ієрархічну складність синтаксичної структури речення і є одним із ключових параметрів у системі автоматичної параметризації [9; 10]. Оповідання демонструють найвищу глибину – 3,21 (σ=0,39), казки – дещо нижчу: 3,11 (σ=0,68), тоді як поезія суттєво відстає: 2,62 (σ=0,56). Відмінність між поезією і оповіданнями становить -18,4%, між поезією і казками – -15,8%, тоді як різниця між двома прозовими жанрами є незначною: лише +3,2%. Вища глибина в оповіданнях відображає більшу ієрархічну вкладеність синтаксичних конструкцій – ланцюги підрядних речень, складні групи іменника з кількома рівнями означень. Нижча глибина в поезії є природним наслідком переважання простих речень і паратаксичних конструкцій, де синтаксична ієрархія є мінімальною. Стандартне відхилення в казках (0,68) майже вдвічі перевищує відповідний показник оповідань (0,39), що відображає жанрову неоднорідність казкової вибірки.

Ширина гілкування – середня кількість безпосередніх залежних на один вузол – є практично ідентичною у казок і оповідань: 4,28 (σ=0,45) і 4,29 (σ=0,26) відповідно, різниця між ними становить лише 0,3%. Поезія помітно відстає: 3,62 (σ=0,54), що на 15,5% менше, ніж у казках, і на 15,8% менше, ніж в оповіданнях. Вищий ступінь гілкування у прозових текстах означає, що кожен головний елемент має в середньому більше безпосередніх підпорядкованих –

розвинені групи іменника з означеннями, дієслівні групи з обставинами і доповненнями. У поезії синтаксичні дерева є структурно «вужчими» і «мілкішими», що цілком узгоджується із загальною синтаксичною стислістю поетичного тексту.

Середня довжина дуги є психолінгвістично значущим параметром: згідно з теорією мінімізації залежнісної відстані [43], чим коротші дуги між головним словом і його залежним, тим легше речення обробляється читачем. Х. Лю встановив, що середня довжина залежності у текстах більшості мов не перевищує трьох слів [43] – наші дані цілком узгоджуються з цією закономірністю. Значення є відносно близькими в усіх трьох жанрах: 2,95 у казках (σ=0,29), 3,06 в оповіданнях (σ=0,20) і 2,83 у поезії (σ=0,30). Відмінність між поезією і оповіданнями становить -7,4%, між поезією і казками – лише -4,1%, між казками і оповіданнями – +3,6%. Оповідання мають дещо більшу середню довжину дуги, що може свідчити про більш аналітичний синтаксис із частішим дистантним розташуванням залежних відносно головного слова. Цей параметр є найменш диференціюючим серед п'яти параметрів дерев залежностей.

Асиметрія графа є єдиним параметром дерев залежностей, за яким поезія показує вище, а не нижче значення порівняно з прозою: 0,51 (σ=0,09) проти 0,44 (σ=0,04) у казках і 0,44 (σ=0,02) в оповіданнях. Відмінність між поезією і казками становить +14,6%, між поезією і оповіданнями – +15,7%, тоді як казки й оповідання є практично ідентичними (-0,9%). Вища асиметрія в поезії вказує на більш виражену лівобічну або правобічну орієнтацію синтаксичних дерев, що пов'язане з особливостями поетичного порядку слів – інверсіями і специфічним розташуванням означень та обставин, зумовленими ритмічними вимогами вірша. Стандартне відхилення у поезії (0,09) є вчетверо більшим, ніж в оповіданнях (0,02), що підтверджує більшу стилістичну варіативність у межах поетичної вибірки. Ящикові діаграми (рис. 3.5) наочно демонструють, що міжквартильний розмах поезії за цим параметром є значно ширшим, ніж у прозових жанрів.

Частка термінальних вузлів є найбільш стабільним параметром серед усіх п'ятнадцяти: 0,66 у казках (σ=0,02) і оповіданнях (σ=0,01) та 0,64 у поезії (σ=0,04). Жодна з міжжанрових відмінностей не перевищує 3,1%. Ця стабільність вказує на те, що незалежно від жанру приблизно дві третини слів у реченні є термінальними – тобто не мають власних залежних. Це є лінгвістичною властивістю синтаксичних дерев природної мови, що підтверджується і міжмовними корпусними дослідженнями [62], і не залежить від жанрової приналежності тексту.

Пунктуаційні індикатори відображають частоту вживання ком, тире та крапок, нормалізовану до 1000 слів тексту. Зведені значення наведено у Таблиці 3.4.

Таблиця 3.4. Пунктуаційні індикатори жанрових вибірок (на 1000 слів, M ± σ)

Параметр	**Казки**	**Оповідання**	**Поезія**
Коми	115,44± 25,24	121,43 ± 19,14	177,22 ± 74,92
Тире	60,71 ± 33,47	42,03 ± 18,93	41,89 ± 27,76
Крапки	77,90 ± 24,54	75,11 ± 16,43	73,74 ± 28,00

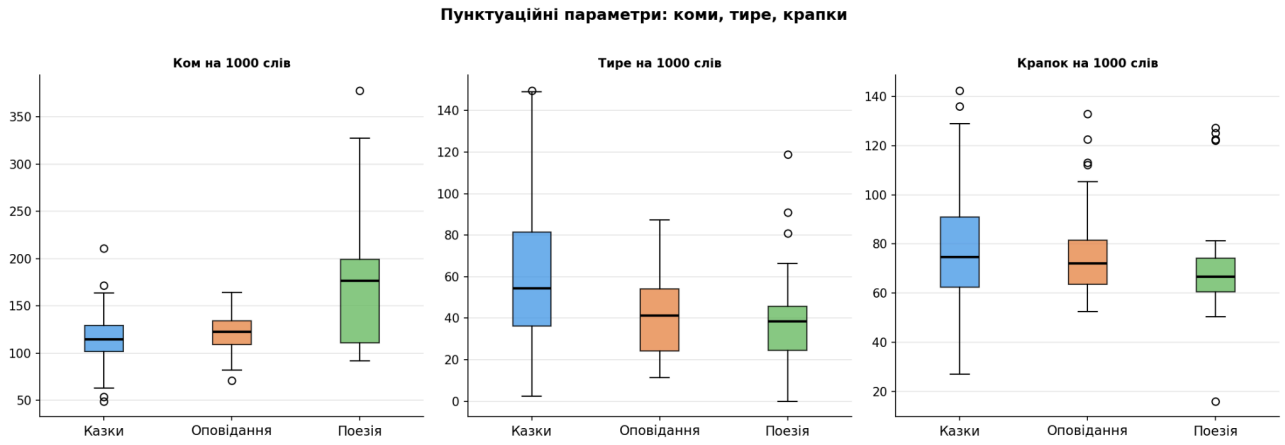

Рис. 3.6. Пунктуаційні параметри: коми, тире, крапки

Частота ком є найбільш диференціюючим пунктуаційним параметром. Поезія демонструє різко вищу щільність ком – 177,22 (σ=74,92) – порівняно з казками (115,44, σ=25,24) і оповіданнями (121,43, σ=19,14). Відмінність між поезією і казками становить +53,5%, між поезією і оповіданнями – +45,9%. Казки й оповідання є близькими за цим параметром: різниця між ними лише 5,2%. Надзвичайно велике стандартне відхилення у поезії (74,92 – майже втричі більше, ніж у казках) підтверджує високу внутрішню варіативність: деякі поетичні тексти є надзвичайно насиченими комами, тоді як інші – зокрема колискові з простою синтаксичною структурою – містять їх значно менше. Висока щільність ком у поезії пояснюється специфікою поетичного синтаксису: короткі рядки часто завершуються комою, що позначає паузу або незавершеність конструкції, а перерахування і звертання, типові для дитячої поезії, також генерують високу щільність ком.

Частота тире є параметром, за яким казки суттєво виділяються на тлі двох інших жанрів: 60,71 (σ=33,47) проти 42,03 (σ=18,93) в оповіданнях і 41,89 (σ=27,76) у поезії. Казки мають тире на 30,8% більше, ніж оповідання, і на 31,0% більше, ніж поезія. Натомість різниця між оповіданнями і поезією є мізерною – лише 0,3%. Висока частота тире в казках пояснюється їхньою діалогічністю: пряма мова персонажів, введена через тире, є одним із ключових стилістичних прийомів казкового тексту [46; 47]. Велике стандартне відхилення у казках (33,47) відображає різний ступінь діалогічності окремих текстів – від мінімально діалогічних описових казок до насичених прямою мовою діалогічних.

Частота крапок є найбільш стабільним пунктуаційним параметром: 77,90 у казках (σ=24,54), 75,11 в оповіданнях (σ=16,43) і 73,74 у поезії (σ=28,00). Жодна з міжжанрових відмінностей не перевищує 5,3%. Ця стабільність є очікуваним результатом: крапка позначає кінець речення, і після нормалізації до 1000 слів кількість крапок є пропорційною до кількості речень на одиницю тексту, яка є відносно сталою в усіх жанрах. Незначна тенденція до зниження від казок до поезії узгоджується з даними про середню довжину речення: чим

довші речення, тим менше їх міститься на 1000 слів і тим менше крапок. Стандартне відхилення у поезії (28,00) є найвищим серед трьох жанрів, що відображає різний обсяг речень у різних поетичних піджанрах.

Зведена теплова карта нормалізованих значень (рис. 3.7) відображає синтаксичні профілі трьох жанрів одночасно і робить очевидною головну закономірність: поезія утворює виразно відокремлений профіль, тоді як казки й оповідання є синтаксично дуже схожими і важко розрізнюваними за кількісними показниками. Найбільш стилерозрізнювальними параметрами виявилися вага іменників (+32–36% у поезії), вага сполучників (-37–40% у поезії), частка простих речень (+60% у поезії відносно прози), частка односкладних речень (50% у поезії проти 20–21% у прозі), а також частота ком (+46–54% у поезії) і тире (+31% у казках відносно двох інших жанрів). Детальна лінгвістична інтерпретація цих закономірностей подана в підрозділі 3.5.

Рис. 3.7. Синтаксичний профіль жанрів (нормалізовані значення)

3.4. Порівняльний аналіз жанрів (3 вибірки × 3 жанри). Стилеметричний аналіз та Візуалізація результатів (діаграми, boxplots, heatmaps).

Порівняльний аналіз трьох жанрових вибірок – казок, оповідань і поезії – здійснювався на основі п'ятнадцяти синтаксичних і морфологічних параметрів, обчислених у підрозділі 3.3. Результати аналізу дозволяють виявити два

принципово різних типи міжжанрових відносин: по-перше, чітке протиставлення поезії обом прозовим жанрам за більшістю параметрів, і по-друге, синтаксичну близькість казок і оповідань, які демонструють статистично подібні профілі за переважною більшістю показників.

Найзручнішим інструментом для узагальненого порівняння є теплова карта нормалізованих значень усіх п'ятнадцяти параметрів (рис. 3.7). На ній кожен параметр нормалізовано до діапазону від 0 до 1 відносно мінімального і максимального значення серед трьох жанрів, що дозволяє порівнювати параметри різних шкал в єдиній системі координат. Теплова карта наочно підтверджує основну закономірність дослідження: рядки «Казки» і «Оповідання» є значно подібнішими між собою, ніж кожен із них до рядка «Поезія». Поезія стабільно демонструє найвищі значення за вагою іменників і частотою ком та найнижчі – за середньою довжиною речення, вагою сполучників, кількістю клаузів, глибиною дерева і шириною гілкування. Казки вирізняються з-поміж трьох жанрів лише за одним параметром – частотою тире, за якою вони помітно перевищують обидва інші жанри.

Порівняльний аналіз морфологічних параметрів засвідчує, що поезія формує якісно відмінний частиномовний профіль. Три з шести морфологічних параметрів виявляють статистично значущі відмінності між поезією і прозою: вага іменників (поезія перевищує казки на 32,2% і оповідання на 36,1%), вага сполучників (поезія нижча за казки на 36,7% і за оповідання на 39,9%) і середня довжина речення (поезія коротша за казки на 22,4% і за оповідання на 26,3%). Натомість вага прикметників, дієслів і прийменників не відображає виражених міжжанрових відмінностей, що відображає їхню меншу стилерозрізнювальну силу в контексті жанрової диференціації дитячої літератури. Казки й оповідання є практично нерозрізненними за всіма морфологічними параметрами: жодна з попарних відмінностей між ними не перевищує 8,1%, тоді як відмінності між поезією і прозою сягають 40%. Ящикові діаграми (рис. 3.1 і 3.2) підтверджують цю закономірність: ящики казок і оповідань значно перекриваються між собою, тоді як ящик поезії за ключовими параметрами є чітко відокремленим.

Порівняльний аналіз структури речень є найбільш показовим розділом дослідження. Стовпчаста діаграма (рис. 3.3) добре ілюструє, що казки і оповідання є практично ідентичними: частка простих речень у казках становить 34,2%, в оповіданнях – 34,5%, а складних – 65,8% і 65,5% відповідно. Поезія різко відрізняється: 54,8% простих речень і 45,2% складних – тобто у поезії прості речення переважають, тоді як у прозі переважають складні. Ще виразніше протиставлення виявляється у розподілі двоскладних і односкладних речень: у казках і оповіданнях близько 79% речень є двоскладними і лише 20–21% – односкладними, тоді як у поезії спостерігається рівне співвідношення 50 на 50. Таке рівне розщеплення є характерною рисою поетичного синтаксису, де номінативні, безособові та еліптичні конструкції є не стилістичним відхиленням, а жанровою нормою. Середня кількість клаузів також підтверджує цю тенденцію: казки (2,46) і оповідання (2,44) є практично однаковими, тоді як поезія має нижче значення (2,10), що на 14–15% менше.

Порівняльний аналіз параметрів дерев залежностей показує, що поезія вирізняється синтаксично «спрощеними» деревами за більшістю показників, тоді як казки й оповідання демонструють близькі значення. Ящикові діаграми (рис. 3.4 і 3.5) наочно відображають цю закономірність: за глибиною дерева, шириною гілкування і довжиною дуги ящики казок і оповідань значно перекриваються, тоді як ящик поезії розташований нижче. Виняток становить асиметрія графа, за якою поезія показує вище значення (0,51), ніж казки (0,44) і оповідання (0,44). Цей єдиний параметр, за яким поезія перевищує прозові жанри, відображає специфіку поетичного порядку слів: інверсії та нестандартне розташування синтаксичних елементів, зумовлені ритмічними і стилістичними вимогами вірша, спричиняють вищу асиметрію залежнісних дерев. Між казками і оповіданнями відмінності за параметрами дерев залежностей є мінімальними: найбільша з них – різниця в глибині дерева (3,11 у казок проти 3,21 в оповіданнях, тобто +3,2%) – є ледь відчутною і не утворює виразного жанрового контрасту.

Порівняльний аналіз пунктуаційних індикаторів виявляє цікаву асиметрію: якщо за морфологічними і синтаксичними параметрами поезія стабільно поступається прозі або перевершує її за іменниковою щільністю, то пунктуаційні параметри дають складнішу картину. За частотою ком поезія різко перевершує обидва прозові жанри (+53,5% відносно казок і +45,9% відносно оповідань), тоді як за частотою тире казки виділяються з-поміж усіх трьох жанрів (+30,8% відносно оповідань і +31,0% відносно поезії). Частота крапок є практично однаковою в усіх трьох жанрах (різниця не перевищує 5,3%). Таким чином, пунктуаційні параметри доповнюють синтаксичну картину: висока щільність ком у поезії підтверджує її паузальну, ритмічно організовану синтаксичну структуру, а висока щільність тире в казках підтверджує їхній діалогічний характер.

Порівняльний аналіз дає підстави для таких висновків. Поезія утворює якісно відмінний синтаксичний профіль: коротші речення, вища іменникова щільність, менше сполучників, більше простих і односкладних речень, менш глибокі й вужчі дерева залежностей, вища асиметрія графа і значно густіші коми – усі ці ознаки взаємопов'язані й разом відображають специфіку поетичного синтаксису. Казки й оповідання, навпаки, виявилися синтаксично дуже схожими: за більшістю параметрів різниця між ними настільки мала, що кількісна стилеметрія на рівні синтаксису не дозволяє їх чітко розмежувати. Єдиний виразний маркер відмінності – частота тире, яка відображає діалогічний характер казкового тексту. Детальну лінгвістичну інтерпретацію цих результатів подано в підрозділі 3.5.

3.5. Лінгвістична інтерпретація параметрів

Отримані кількісні дані потребують лінгвістичного пояснення – потрібно встановити зв'язок між числовими показниками і жанровими властивостями досліджуваних текстів. Ми зосереджуємося на трьох питаннях: чому поезія формує окремий синтаксичний профіль, чому казки й оповідання синтаксично близькі і що саме дозволяє їх все ж розрізнити.

Синтаксичний профіль поезії є найбільш виразним і лінгвістично обґрунтованим результатом дослідження. Сукупність параметрів, за якими поезія відрізняється від прози, утворює внутрішньо узгоджену картину, яка відображає фундаментальні жанрові властивості поетичного тексту. Коротші речення (8,22 слова проти 10–11 у прозі), вища частка простих речень (54,8% проти 34%), рівне співвідношення двоскладних і односкладних речень (50/50 проти 79/21 у прозі), менша кількість клаузів (2,10 проти 2,44–2,46), менша глибина дерева залежностей (2,62 проти 3,11–3,21) і менша ширина гілкування (3,62 проти 4,28–4,29) – усі ці показники вказують на одне: поетичний текст організовується принципово іншими синтаксичними засобами, ніж прозовий. Ця відмінність є не випадковою, а зумовленою самою природою поетичного мовлення, в якому ритм, метр і строфічна будова виступають структуруючими принципами, що замінюють або доповнюють граматичний синтаксис.

Висока частка іменників у поезії (30,38% проти 22–23% у прозі) є найбільш показовим результатом. Ця закономірність пояснюється тим, що дитяча поезія будується переважно на образах-предметах і явищах: лексика колискових, лічилок, скоромовок, дитячих віршів є переважно конкретною, предметною, іменниковою за своєю природою. Крім того, висока частка іменників корелює з поширеністю в поетичних текстах номінативних речень – конструкцій, де іменник виступає єдиним головним членом. Саме ці конструкції становлять значну частину односкладних речень, зафіксованих у поетичній вибірці, і забезпечують ефект синтаксичної стислості та образної насиченості, характерний для поетичного мовлення. Дж. Піаже показав, що на доопераційній стадії розвитку (2–7 років) дитина мислить переважно конкретними, предметними категоріями [27], і дитяча поезія, зорієнтована на цю вікову групу, природно відображає цю особливість когнітивного розвитку у своєму лексичному і синтаксичному устрої. Статистичні параметри функціональних стилів досліджувалися С. Бук [66] та В. Перебийніс [67].

Різко нижча частка сполучників у поезії (5,55% проти 8,77–9,24% у прозі) є прямим синтаксичним відображенням жанрової стратегії паратаксису.

Паратаксис – розташування синтаксично незалежних або слабко пов'язаних конструкцій без граматичних сполучників – є одним із ключових прийомів поетичного синтаксису. У дитячій поезії він реалізується через переліки, звертання, вигуки і фрагментарні конструкції, що розташовуються в межах рядка або строфи без сурядних чи підрядних сполучників. Це пояснює також і нижчу частку складних речень у поезії (45,2% проти 65–66% у прозі): складні речення передбачають граматично оформлений зв'язок між клаузами, а в паратаксичному тексті цей зв'язок замінюється ритмічним і семантичним.

Підвищена асиметрія синтаксичних дерев у поезії (0,51 проти 0,44 у прозі) є ще одним лінгвістично значущим результатом. Асиметрія графа відображає переважання залежностей одного напряму – лівих або правих – у структурі дерева. У поетичному тексті ця асиметрія зростає через інверсії і нестандартний порядок слів, зумовлений ритмічними й евфонічними вимогами: означення може стояти після означуваного, обставина – перед дієсловом, підмет – у кінці речення. Такий порядок слів є поетичною нормою, а не відхиленням, і природно породжує вищу асиметрію залежнісних дерев. Це спостереження узгоджується з результатами Н. Дарчук і В. Сорокіна [9; 10], які встановили, що асиметрія графа є одним із параметрів, що мають стилерозрізнювальну силу для поетичного мовлення.

Висока частота ком у поезії (177,22 на 1000 слів проти 115–121 у прозі) є пунктуаційним відображенням ритмічно-паузальної організації поетичного тексту. У дитячому вірші кома часто позначає не стільки граматичну межу, скільки паузу для дихання або ритмічний перебій, що є невіддільним елементом поетичного звучання. Перерахування, типові для дитячої поезії (назви тварин, явищ, предметів), також генерують високу щільність ком. Великий розкид цього параметра в поетичній вибірці (σ=74,92) відображає жанрову різнорідність групи: колискові і ліричні вірші мають одну пунктуаційну логіку, тоді як лічилки, скоромовки і жниварські пісні – зовсім іншу.

Синтаксична близькість казок і оповідань є другим ключовим результатом дослідження і потребує окремого лінгвістичного пояснення. Незважаючи на

очевидні жанрові відмінності – казки є фантастичним наративом з усталеною структурою, оповідання – реалістичним прозовим текстом з різноманітною тематикою – обидва жанри демонструють практично ідентичні синтаксичні профілі за переважною більшістю параметрів. Ця подібність не випадкова і пояснюється кількома факторами. По-перше, обидва жанри є прозовими наративами, орієнтованими на ту саму вікову групу читачів – дошкільний і молодший шкільний вік, – що зумовлює схожий рівень синтаксичної складності. По-друге, і казки, і оповідання реалізують основну наративну стратегію – послідовний опис подій і вчинків, яка передбачає переважання складних речень із причиново-наслідковим зв'язком, розгорнутих груп дієслова і підрядних конструкцій. По-третє, В. Пропп переконливо показав, що казка як жанр є глибоко структурованим наративом із сталою послідовністю функцій [46], що породжує типові синтаксичні конструкції – послідовні дієслівні ланцюги, прямолінійна синтаксична будова, чіткі причиново-наслідкові відносини. Ці конструкції є за своєю природою схожими до тих, що характеризують реалістичне оповідання, хоча й реалізуються на іншому тематичному і стилістичному матеріалі.

Єдиним параметром, що виразно диференціює казки від оповідань, є частота тире (60,71 у казках проти 42,03 в оповіданнях – різниця +44,4% на користь казок). Цей параметр є непрямим маркером діалогічності – частки прямої мови в тексті. Висока щільність тире в казках відображає їхню фундаментальну жанрову рису: казковий текст є значною мірою діалогічним, побудованим на прямій мові персонажів, введеній через тире. Л. Дунаєвська підкреслює, що українська народна казка характеризується специфічною ритмо-синтаксичною організацією, де діалог відіграє ключову структуроутворюючу роль [47]. Оповідання, особливо реалістичні твори для молодшого шкільного віку, також містять діалоги, однак значно менш насичені ними. Таким чином, тире є єдиним із п'ятнадцяти досліджуваних параметрів, що дозволяє кількісно зафіксувати відому якісну відмінність між казкою і оповіданням на рівні діалогічності тексту.

Слід окремо зупинитися на питанні про обмеження лінгвістичної інтерпретації результатів. По-перше, мала чисельність поетичної вибірки (21 текст) означає, що виявлені закономірності мають попередній характер і потребують верифікації на більшому масиві поетичних текстів. По-друге, жанрова неоднорідність поетичної вибірки – яка включає вірші, колискові, лічилки, скоромовки і народні пісні – може маскувати внутрішньожанрові відмінності між цими піджанрами: цілком можливо, що лічилки і скоромовки мають відмінний синтаксичний профіль порівняно з ліричними віршами, однак через малу кількість текстів ці відмінності усереднюються. По-третє, висока внутрішньожанрова варіативність казок (що відображається у великих стандартних відхиленнях за більшістю параметрів) показує, що казкова вибірка є жанрово і стилістично неоднорідною: народні казки-анекдоти, авторські казкові повісті і казки-п'єси мають різні синтаксичні профілі, які усереднюються в загальному показнику. Попри ці обмеження, виявлені закономірності є лінгвістично обґрунтованими і узгоджуються як із теоретичними уявленнями про жанрову специфіку досліджуваних текстів, так і з даними аналогічних зарубіжних досліджень [13; 16].

Отже, автоматична синтаксична параметризація підтвердила вихідну гіпотезу: жанрова приналежність тексту є стилеметрично значущим фактором і відображається у кількісних синтаксичних параметрах. Поезія утворює виразно відокремлений профіль – через ритмічну організацію, паратаксис, номінальну щільність і строфічну будову. Казки й оповідання є синтаксично подібними прозовими жанрами, і єдиний параметр, що дає змогу їх відрізнити, – це частота тире як показник діалогічності. Ці результати відкривають напрями для подальшої роботи: порівняльний аналіз окремих піджанрів і верифікація закономірностей на ширшому і збалансованішому корпусі.

ВИСНОВКИ ДО РОЗДІЛУ 3

У цьому розділі ми здійснили автоматичну синтаксичну параметризацію корпусу українськомовної дитячої літератури і провели порівняльний стилеметричний аналіз трьох жанрових вибірок.

Ми сформували три вибірки – 97 казок, 57 оповідань і 21 поетичний текст (загалом 70 287 речень) – і опрацювали їх скриптом stanza_analysis.py, що реалізує чотирьохетапний пайплайн Stanza. Параметризацію ми здійснювали за 15 кількісними показниками, згрупованими у чотири блоки: морфологічні параметри, структура речень, параметри дерев залежностей і пунктуаційні індикатори.

Порівняльний аналіз виявив два типи міжжанрових відносин. Перший – чітке протиставлення поезії і прози: поезія відрізняється від казок і оповідань за більшістю параметрів, і ця відмінність є лінгвістично обґрунтованою – ритмічна організація, паратаксис, номінальна щільність і строфічна будова разом формують якісно відмінний синтаксичний профіль. Другий тип – синтаксична близькість казок і оповідань: за більшістю параметрів різниця між ними статистично незначуща. Єдиний параметр, що виразно розрізняє два прозові жанри, – частота тире (60,71 у казках проти 42,03 в оповіданнях), яка є непрямим кількісним маркером діалогічності казкового тексту [47].

ВИСНОВКИ

Дослідження підтвердило, що жанрова приналежність українськомовного дитячого тексту є стилеметрично значущим фактором, який фіксується кількісними синтаксичними параметрами. Отримані результати дозволяють вважати всі дев'ять поставлених завдань виконаними.

Опрацювання наукової літератури засвідчило, що синтаксична стилеметрія є найменш дослідженою ланкою української стилеметричної традиції, а дитяча література – перспективним, але практично недослідженим об'єктом кількісного аналізу в Україні.

Структурування корпусу та формування трьох жанрових вибірок (97 казок, 57 оповідань, 21 поетичний текст — загалом 175 текстів, 70 287 речень, 741 780 слововживань) забезпечило матеріальну базу для подальшого аналізу. Порівняльна апробація трьох NLP-інструментів (Stanza, spaCy, UDPipe) виявила їхні жанрово зумовлені слабкі місця і підтвердила, що Stanza забезпечує найстабільнішу сегментацію речень на матеріалі всіх трьох жанрів.

Автоматична параметризація за 15 кількісними показниками і порівняльний жанровий аналіз виявили два типи міжжанрових відносин. Поезія утворює якісно відмінний синтаксичний профіль – коротші речення, вища іменникова щільність, менше сполучників, більше простих і односкладних конструкцій, паратаксис і ритмічна пунктуація. Казки й оповідання є синтаксично близькими прозовими жанрами, і єдиним параметром, що їх розрізняє, є частота тире як кількісний маркер діалогічності.

Розроблене програмне забезпечення – скрипти stanza_analysis.py і visualize_and_save.py – забезпечує автоматичний синтаксичний аналіз і розрахунок стилеметричних параметрів для будь-якого українськомовного корпусу і є практичним результатом дослідження.

Наукова новизна роботи полягає у першій спробі комплексної синтаксичної стилеметрії українськомовної дитячої літератури із залученням автоматизованого парсингу. Практична цінність полягає у можливості застосування розробленого інструментарію для автоматизованої оцінки складності дитячих текстів, їхнього жанрового розмежування та укладання навчальних матеріалів.

Для лінгвостилістики значення роботи полягає в тому, що це перше дослідження, яке застосувало комплексну синтаксичну параметризацію до українськомовної дитячої літератури. Параметри дерев залежностей, а саме глибина, ширина гілкування, асиметрія графа виявилися надійними жанровими індикаторами, що розширює методологічний арсенал української стилеметрії за рахунок синтаксичного рівня.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Stamatatos E. A Survey of Modern Authorship Attribution Methods. URL: https://icsdweb.aegean.gr/stamatatos/papers/survey.pdf

2. Перебийніс В. С. Частотний словник сучасної української художньої прози в двох томах. Київ: Наукова думка, 1981.

3. Перебийніс В. І. Статистичні методи для лінгвістів: Навч. посібник. Вінниця: Нова книга, 2002.

4. Карпіловська Є. А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика. Донецьк, 2006.

5. Бук С. Статистичні характеристики лексики основних функціональних стилів української мови. 2006. URL: https://nasplib.isofts.kiev.ua/items/78914193-e608-4065-bef6-5acf748ee6f0

6. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology. Eastern-European Journal of Enterprise Technologies. 2017. № 4(2). С. 10–19. URL: http://nbuv.gov.ua/UJRN/Vejpte_2017_4(2)__3

7. Бісікало О. В., Іщенко О. Р. Визначення авторства україномовного тексту на основі методів машинного навчання. URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/48024/24725.pdf

8. Darchuk N., Robeiko V., Zuban O. The System for Automatic Stylometric Analysis of Ukrainian Media Texts TextAttributor 1.0. 2024. URL: https://www.researchgate.net/publication/387309091

9. Darchuk N., Sorokin V. Parameterization of the Ukrainian Text Corpus Based on Parsing Results. URL: https://ceur-ws.org/Vol-3171/paper23.pdf

10. Darchuk N., Zuban O., Sorokin V. On Stylistic Differentiation in Ukrainian Poetic Speech Based on the Syntactic Structure of Sentences. 2024. URL: https://www.researchgate.net/publication/380478965

11. Дарчук Н. П., Цигвінцева Ю. О. Комп'ютерна параметризація українськомовного тексту: навчальний посібник. Київ: КНУ імені Тараса Шевченка, 2025. 180 с.

12. Гільтайчук В. Автоматичне укладання частотного словника української мови як засобу мовного розвитку дітей дошкільного віку. URL: https://ir.library.knu.ua/entities/publication/6bfa1be1-0c9e-4d5f-ba67-4a6a5c164387

13. Haverals W., Geybels L., Joosen V. A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults. Language and Literature: International Journal of Stylistics. 2022. URL: https://www.researchgate.net/publication/357919736_A_style_for_every_age_A_stylometric_inquiry_into_crosswriters_for_children_adolescents_and_adults

14. Жуковська В. В. Вступ до корпусної лінгвістики. URL: https://eprints.zu.edu.ua/18909/1/korpusna_lingv.pdf

15. Дарчук Н. П. Можливості семантичної розмітки Корпусу української мови (КУМ). URL: https://enpuirb.udu.edu.ua/server/api/core/bitstreams/2c1a7749-0bf0-4331-98b2-b4f8f4d65dfc/content

16. Hsiao Y., Dawson N. J., Banerji N., Nation K. A corpus-based developmental investigation of linguistic complexity in children's writing. Applied Corpus Linguistics. 2024. Vol. 4, No. 1. Article 100084. URL: https://www.sciencedirect.com/science/article/pii/S2666799124000017

17. Gorrell P. Syntax and Parsing. Cambridge: Cambridge University Press, 1995.URL:https://books.google.com.ua/books?id=T45nULUlJqwC&printsec=copyright&redir_esc=y

18. Marie-Catherine de Marneffe M.-C., Manning C., Nivre J., Zeman D. Universal Dependencies. Computational Linguistics. 2021. Vol. 47, No. 2. P. 255–308. URL: https://aclanthology.org/2021.cl-2.11/

19. Kotsyba N., Moskalevskyi B., Romanenko M. UD Ukrainian-IU. Universal Dependencies, 2016–2025. URL: https://universaldependencies.org/treebanks/uk_iu/index.html

20. Straka M., Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task. Vancouver, 2017. P. 88–99. URL: https://ufal.mff.cuni.cz/udpipe/2

21. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Proceedings of ACL 2020: System Demonstrations. P. 101–108. URL: https://stanfordnlp.github.io/stanza/

22. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. AIST 2015. Communications in Computer and Information Science. Vol. 542. Springer, 2015. P. 320–332. DOI: 10.1007/978-3-319-26123-2_31. URL: https://pymorphy2.readthedocs.io/

23. Shvedova M., Lukashevskyi A., Rysin A. Developing a Universal Dependencies Treebank for Ukrainian Parliamentary Speech. Proceedings of UNLP 2025. P. 55–63. URL: https://aclanthology.org/2025.unlp-1.7/

24. Honnibal M., Montani I., Van Landeghem S., Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo, 2020. DOI: 10.5281/zenodo.1212303. URL: https://spacy.io/

25. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. draft. Stanford University, 2023. URL: https://web.stanford.edu/~jurafsky/slp3/

26. Kübler S., McDonald R., Nivre J. Dependency Parsing. Synthesis Lectures on Human Language Technologies. Vol. 2. Morgan & Claypool Publishers, 2009. 127 p. DOI: 10.2200/S00169ED1V01Y200901HLT002

27. Piaget J. The Psychology of Intelligence. London: Routledge & Kegan Paul, 1950. 182 p.

28. Lété B., Sprenger-Charolles L., Colé P. MANULEX: A grade-level lexical database from French elementary school readers. Behavior Research Methods, Instruments, & Computers. 2004. Vol. 36, No. 1. P. 156–166. URL: https://link.springer.com/article/10.3758/BF03195560

29. Schroeder S., Würzner K.-M., Heister J., Geyken A., Kliegl R. childLex: a lexical database of German read by children. Behavior Research Methods. 2015. Vol. 47, No. 4. P. 1085–1094. URL: https://pubmed.ncbi.nlm.nih.gov/25319039/

30. Шведова М., фон Вальденфельс Р., Яригін С., Рисін А., Старко В., Ніколаєнко Т. та ін. Генеральний регіонально анотований корпус української мови (ГРАК). Київ, Львів, Єна, 2017–2024. URL: https://uacorpus.org

31. Oxford Wordlist. Oxford University Press Australia, 2024. URL: https://www.oxfordwordlist.com/

32. El Corpus de Referencia del Español Actual (CREA). Real Academia Española. URL: https://www.rae.es/banco-de-datos/crea

33. Burrows J. 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing. 2002. Vol. 17, No. 3. P. 267–287. URL: https://www.researchgate.net/publication/240956478_'Delta'_A_Measure_of_Stylistic_Difference_and_a_Guide_to_Likely_Authorship

34. Koppel M., Schler J., Argamon S. Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology. 2009. Vol. 60, No. 1. P. 9–26. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961

35. Mosteller F., Wallace D. L. Inference in an Authorship Problem. Journal of the American Statistical Association. 1963. Vol. 58, No. 302. P. 275–309. URL: https://www.jstor.org/stable/2283270

36. Eder M., Rybicki J., Kestemont M. Stylometry with R: A Package for Computational Text Analysis. R Journal. 2016. Vol. 8, No. 1. P. 107–121. URL: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

37. Holmes D. I. The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing. 1998. Vol. 13, No. 3. P. 111–117. URL: https://academic.oup.com/dsh/article-abstract/13/3/111/933269

38. McEnery T., Wilson A. Corpus Linguistics: An Introduction. 2nd ed. Edinburgh: Edinburgh University Press, 2001. 235 p. URL:

https://www.academia.edu/1171341/McEnery_T_and_Wilson_A_2001_Corpus_Linguistics_second_edition_

39. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998. 300 p. URL: https://books.google.com.ua/books/about/Corpus_Linguistics.html?id=2h5F7TXa6psC&redir_esc=y

40. Kennedy G. An Introduction to Corpus Linguistics. London: Addison Wesley Longman, 1998. 315 p. URL: https://scispace.com/papers/an-introduction-to-corpus-linguistics-2hfltbd2y6

41. Sinclair J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991. 179 p. URL: https://www.academia.edu/16757872/SInclair_Corpus_Concordance_Collocation

42. Leech G. Corpora and theories of linguistic performance. Directions in Corpus Linguistics. Berlin: Mouton de Gruyter, 1992. P. 105–122. URL: https://www.scirp.org/reference/referencespapers?referenceid=2734035

43. Liu H. Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science. 2008. Vol. 9, No. 2. P. 159–191. URL: http://www.lingviko.net/JCS.pdf

44. Nivre J. Dependency Grammar and Dependency Parsing. MSI report 05133. Växjö University, 2005. 32 p. URL: https://cl.lingfil.uu.se/~nivre/docs/05133.pdf

45. Straka M., Hajič J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files. Proceedings of LREC 2016. Portorož, 2016. P. 4290–4297.

46. Propp V. Morphology of the Folktale. 2nd ed. Austin: University of Texas Press, 1968. 185 p.

47. Дунаєвська Л. Ф. Українська народна казка. Київ: Вища школа, 1987. 127 с.

48. Богуш А. М. Мовленнєвий розвиток дітей від народження до 7 років. Київ: Слово, 2004. 374 с.

49. Krashen S. The Input Hypothesis: Issues and Implications. London: Longman, 1985. 120 p. URL: https://archive.org/details/inputhypothesisi0000kras/page/n1/mode/2up

50. Hunt K. W. Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3. Champaign: National Council of Teachers of English, 1965. 184 p. URL: https://files.eric.ed.gov/fulltext/ED113735.pdf

51. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press, 1999. 680 p.

52. Oakes M. P. Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press, 1998. 287 p. URL: https://coehuman.uodiyala.edu.iq/uploads/Coehuman%20library%20pdf/English%20library%D9%83%D8%AA%D8%A8%20%D8%A7%D9%84%D8%A7%D9%86%D9%83%D9%84%D9%8A%D8%B2%D9%8A/linguistics/statistics.pdf

53. Gries S. T. Quantitative Corpus Linguistics with R: A Practical Introduction. New York: Routledge, 2009. 238 p. URL: https://www.stgries.info/research/2016_STG_QCLWR2.pdf

54. Kilgarriff A. Comparing Corpora. International Journal of Corpus Linguistics. 2001. Vol. 6, No. 1. P. 97–133. URL: https://www.researchgate.net/publication/263252975_Comparing_Corpora

55. Єрмоленко С. Я. Нариси з української словесності: стилістика та культура мови. Київ: Довіра, 1999. 431 с.

56. Пономарів О. Д. Стилістика сучасної української мови. 3-тє вид. Тернопіль: Навчальна книга – Богдан, 2000. 248 с.

57. Чабаненко В. А. Стилістика експресивних засобів української мови. Запоріжжя: ЗДУ, 2002. 351 с.

58. Кочан І. М. Лінгвістичний аналіз тексту. Київ: Знання, 2008. 423 с.

59. Загнітко А. П. Теоретична граматика сучасної української мови. Синтаксис. Донецьк: ДонНУ, 2001. 662 с.

60. Zeman D. et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Proceedings of CoNLL 2017. 2017. P. 1–19. URL: https://aclanthology.org/K17-3001/

61. Qi P. et al. Universal Dependency Parsing from Scratch. Proceedings of CoNLL 2018 Shared Task. 2018. P. 160–170. URL: https://aclanthology.org/K18-2016/

62. Nivre J. et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. Proceedings of LREC 2020. 2020. URL: https://aclanthology.org/2020.lrec-1.497/

63. Flesch R. A New Readability Yardstick. Journal of Applied Psychology. 1948. Vol. 32, No. 3. P. 221–233.

64. Crossley S. A., Allen D., McNamara D. S. Text Simplification and Comprehensible Input. Language Teaching Research. 2012. Vol. 16, No. 3. P. 391–403. DOI: 10.1177/1362168811423456

65. Nation I. S. P. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press, 2001. 477 p.

66. Бук С. Статистичні параметри стилів (на матеріалі прози Івана Франка). Вісник Львівського університету. Серія філологічна. 2008. Вип. 44. С. 263–270.

67. Перебийніс В. С. Статистичні параметри стилів. Київ: Наукова думка, 1967. 264 с.

68. Langer J. A. Children Reading and Writing: Structures and Strategies. Norwood: Ablex, 1986. 193 p.

69. Tesnière L. Éléments de syntaxe structurale. Paris: Klincksieck, 1959. 670 p.

70. Halliday M. A. K. An Introduction to Functional Grammar. London: Edward Arnold, 1985. 387 p.

71. Вихованець І. Р. Граматика української мови. Синтаксис. Київ: Либідь, 1993. 368 с.

72. Stubbs M. Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture. Oxford: Blackwell, 1996. 267 p.

73. Covington M. A. A Fundamental Algorithm for Dependency Parsing. Proceedings of the 39th Annual ACM Southeast Conference. Athens, 2001. P. 95–102. URL: https://ai1.ai.uga.edu/mc/dparser/dgpacmnew.pdf

74. Tomasello M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge: Harvard University Press, 2003. 388 p.

75. Čermák F. Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University Press, 2009. 288 p. URL: https://press.umich.edu/pdf/9780472033850-part1.pdf

76. Leech G., Rayson P., Wilson A. Word Frequencies in Written and Spoken English: Based on the British National Corpus. London: Longman, 2001. 320 p.

77. Ninio A., Snow C. E. Pragmatic Development. Boulder: Westview Press, 1996. 230 p. URL: https://www.researchgate.net/publication/273093099_Pragmatic_Development

78. Загнітко А. П. Теорія сучасного синтаксису. Донецьк: ДонНУ, 2006. 378 с.

79. Пентилюк М. І. Культура мови і стилістика. Київ: Вежа, 1994. 240 с.

80. Стишов О. А. Динамічні процеси в лексико-семантичній системі та в словотворі української мови кінця XX ст. Київ: Пугач, 2003. 261 с.

81. Левицький В. В. Квантитативні методи в лінгвістиці. Вінниця : Нова Книга, 2007. 264 с.

82. Перебийніс В. С., Муравицька М. П., Дарчук Н. П. Частотні словники та їх використання. Київ, 1985.

ДОДАТОК 1

Доступ до папки з програмним забезпеченням знаходиться за посиланням: https://drive.google.com/drive/folders/1H5g_3YuNYnTi15-2JAwhhoc4dzVTXvup?usp=sharing