Міністерство освіти і науки України
Український державний університет науки і технологій

Кваліфікаційна наукова
праця на правах рукопису

Демидович Інна Миколаївна

УДК 004.048+004.912

ДИСЕРТАЦІЯ

РОЗВИТОК МЕТОДІВ ТА ЗАСОБІВ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНИХ ТЕКСТІВ НА ОСНОВІ КОНСТРУКТИВНО-ПРОДУКЦІЙНОГО МОДЕЛЮВАННЯ

122 – комп’ютерні науки
12 – інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____І. М. Демидович

Науковий керівник Шинкаренко Віктор Іванович
доктор технічних наук, професор

Дніпро – 2023

АНОТАЦІЯ

Демидович І. М. Визначення авторства природньомовних текстів методами та засобами конструктивно-продукційного моделювання.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» – Український державний університет науки і технологій, Дніпро, 2023.

Дисертація присвячена дослідженню та розробці різних методів й засобів встановлення авторства природньомовних текстів на основі різних показників, що відображають особливості авторського стилю мовлення.

У дисертаційній роботі отримані нові науково обґрунтовані теоретичні та експериментальні результати, що у сукупності дозволять застосовувати досліджені методи самостійно або у комплексі з іншими для встановлення авторства текстів та пошуку запозичень.

У першому розділі виконано огляд та аналіз існуючих наразі методів та підходів, що допомагають вловити авторський стиль для різних мов світу. Показано, що різні підходи зумовлені складністю задачі та особливостями різних мов. Встановлено, що досконалого 100% результату у питанні встановлення авторства текстів досі не набуто, незважаючи на широкий перелік використаних інструментів та підходів.

Виявлено, що дослідження підходів для роботи саме з україномовними текстам мають невеликий відсоток на відміну від робіт присвячених іншим мовам, що зумовлено складністю нормалізування та вільністю побудови речень.

З'ясовано, що через особливості побудови речень українською мовою, широкі можливості автора щодо надання тексту певної стилістики на вимогу ідеї твору чи призначенні роботи, поширені методи та підходи роботи з іншими мовами не зможуть в достатній мірі відобразити авторський стиль.

У другому розділі представлені досліджені методи та розроблені моделі статистичного аналізу, аналізу складності текстів, рекурентного аналізу конструктивно-продукційного моделювання.

Виконано адаптацію методів для роботи з природньомовними текстами українською мови. Запропоновано метод створення профілю автора та метод роботи з багатьма показниками для найкращого врахування особливостей авторського стилю.

Розроблена модель природньомовного тексту у вигляді множини правил стохастичних граматик та розроблені метод порівняння текстів на основі порівняння цих правил, що дозволяє враховувати синтаксичні та стилістичні особливості тексту автора

Розроблені конструктори для перетворення природньомовного тексту на множину стохастичних правил та подальше порівняння таких множин для встановлення ступеня їх співпадіння.

У третьому розділі приведені результати експериментальних досліджень. Перевірена та підтверджена ефективність кожного з методів та розроблених моделей. Виконано експерименти за допомогою репрезентативних вибірків як художніх творів різних авторів, так технічних текстів різного розміру та складу. Встановлено ступінь ефективності кожного з досліджених методів окремо.

В подальшому методи було об’єднано для отримання кращого результату та врахування різних особливостей авторського стилю. Було розвинуто та експериментально доведено ефективність методів роботи з великою кількістю різних показників для отримання кращого результату.

У четвертому розділі розроблено інструменти для автоматичного аналізу тексту, підрахунку відповідних показників та подальшого порівняння робіт за ними. Та інструменти що на основі розроблених конструкторів автоматично будують множини правил для різних текстів та порівнюють обрані на ступінь схожості.

Ключові слова: багатокритеріальна оптимізація, генетичний алгоритм, рекурентний аналіз, розпізнавання образів, конструктивне моделювання, авторство текстів, стохастичні граматики, формальні мови, природньомовні тексти, атрибуція текстів, українська мова, авторська атрибуція, критерій Стьюдента.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Праці у фахових виданнях затверджених МОН України:

1. Shynkarenko, V. I., and Demidovich, I. M. "Determination of the attributes of authorship of natural texts." *Artificial intelligence* 3 (2018): 27-35.
2. Shynkarenko, V. I., Demidovich, I. M., and Kuropiatnyk, O. S. "A Dual Approach to Establishing the Authority of Technical Natural Language Texts and Their Components." *Science and Transport Progress* 2 (102) (2023): 71-85. doi: 10.15802/stp2023/288958.
3. Shynkarenko, V. I., and Demydovych, I. M. "Methods and software for significant indicators determination of the natural language texts author profile." *Problems in programming* 3 (2023): 22-29. doi: 10.15407/pp2023.03.22
4. Shynkarenko, Viktor, and Demidovich, Inna. "Constructive-synthesizing modeling of natural language texts." *Computer systems and information technologies* 32023, p. 81-91. doi: 10.31891/csit-2023-3-10

Праці включені до міжнародних наукометричних баз (МНБД) Scopus та Web of Science:

5. Shynkarenko, Viktor, and Demidovich, Inna. "Natural Language Texts Authorship Establishing Based on the Sentences Structure." *COLINS,* 2022, p. 328-337.
6. Shynkarenko, Viktor I., et al. " Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task." *IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT),* 2021, p. 48-51. doi: 10.1109/CSIT52700.2021.9648829.
7. Shynkarenko, Viktor I., and Demidovich, Inna. "Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights." *5th International Conference on Computational Linguistics and Intelligent Systems (COLINS),* 2021, p. 832-844.

Матеріали міжнародних наукових конференцій:

8. Шинкаренко, В.І, та Демидович, І.М. . “Статистичний та рекурентний аналіз природньомовних. текстів”. *Збірка тез XII Міжнародної науково-практичної*

конференції "Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті", Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2018, с. 120.

9. Шинкаренко, В.И., та Демидович, И.Н. "Рекуррентный анализ естественно-языковых текстов". *Збірка тез Всеукраїнської науково-методичної конференції «Проблеми математичного моделювання»*. Дніпропетровський державний технічний ун-т, 2018, с. 40-43.

10. Шинкаренко, В.І, та Демидович, І.М. "Використання генетичного алгоритму для покращення визначення авторства природньомовних текстів". *Збірка тез XIV Міжнародної науково-практичної конференції "Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті",* Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2020, с. 127.

11. Шинкаренко, В.І, та Демидович, І.М. "Показатель структурного сходства естественно языкового литературного текста". *Збірка тез XV Міжнародної науково-практичної конференції "Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті",* Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2021, с. 68.

12. Шинкаренко, В.І, та Демидович І.М. . "Застосування формальних стохастичних граматик при визначенні авторству текстів" *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2022, с. 293

13. Шинкаренко, В.І, та Демидович, І.М. «Застосування конструктивного моделювання при визначенні авторства текстів» *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2023, с.394.

ABSTRACT

Demidovych I. M. Methods and tools development for Ukrainian-language texts authorship determining based on constructive-synthesizing modeling.

Thesis submitted for obtaining the Doctor of Philosophy degree in the specialty 122 "Computer Sciences" – Ukrainian State University of Science and Technology, Dnipro, 2023.

The dissertation is devoted to the research and various methods and means development for establishing the natural language texts authorship based on various indicators that reflect the peculiarities of the author's speech style.

New theoretical and experimental scientifically based results were obtained, which together will allow applying the researched methods independently or in combination with others to establish the authorship of texts and search for borrowings.

In the first chapter, a review and analysis of currently existing methods and approaches that help to capture the author's style for different languages of the world is performed. It is shown the variety of different existing approaches due to the complexity of the task and the structure distinction in different languages. It has been established that a perfect 100% result in establishing the texts authorship has not yet been achieved, despite the wide range of tools and approaches used.

It was found that the research of approaches for working specifically with Ukrainian-language texts has a small percentage, in contrast to works devoted to other languages, which is due to the complexity of its formalization and the variety of sentence constructions.

It has been found that due to the complexity of sentence structure in the Ukrainian language, and the wide possibilities for the author to provide the text with a certain style at the request of the main idea or the purpose of the work, commonly used methods and approaches will not be able to sufficiently reflect the author's style.

The second section presents the researched methods and developed models statistical analysis, analysis of text complexity, recurrent analysis, structural and production modeling. Methods adaptation for working with natural language texts in the

Ukrainian language has been developed. An author's profile creating and working with the range of indicators, finding the best among them to reflect the author's style crucial features methods are proposed.

A natural language text model in the form of stochastic grammars rules set was developed and the texts comparing method based on the comparison of these rules was developed, which allows working with the syntactic and stylistic features of the author's text.

Constructors have been developed for converting natural language text into a set of stochastic rules and further comparing such sets to establish the degree of their similarity.

The third section presents the results of experimental research. The effectiveness of each method and developed model has been tested and confirmed. Experiments were carried out with the help of representative samples: different authors fictional works and technical texts in different sizes and formats. The effectiveness degree of each investigated method was determined separately.

The methods were combined to obtain a better result and take into account various features of the author's style. The effectiveness of methods working with a large number of different indicators to obtain a better result was developed and experimentally proven.

In the fourth chapter, tools are developed for automatic text analysis, calculation of relevant indicators and further comparison of works based on them. And tools based on developed constructors that automatically build sets of rules for different texts and compare the selected ones for the degree of similarity.

Keywords: multicriteria optimization, genetic algorithm, recurrent analysis, pattern recognition, constructive-synthesizing modeling, authorship of texts, stochastic grammars, formal languages, natural language texts, attribution of texts, Ukrainian language, authorship attribution, Student's criterion.

Оглавление

ВСТУП

Актуальність теми. Ідентифікація авторства – це практичне завдання пошуку найімовірнішого автора тексту. Однак, можливості сучасних методик визначення авторства текстів різних призначення та стилю, як і раніше, обмежені, а результати не завжди відповідають реальності. Визначення авторства тексту досі є не вирішеною проблемою для багатьох сфер діяльності людини, таких як освіта, юриспруденція, літературознавство, історія тощо.

Проблематика виявлення схожості та розбіжностей в текстах різних авторів є досі актуальною через складності ідентифікації збіжностей, які не є прямим співпадінням тексту. Особливу складність становить робота з особливостями, характерними для конкретної мови, що суттєво ускладнює поставлене завдання та робить неможливим створення уніфікованого інструментарію.

Методику, що дозволяє з достатньою достовірністю визначати авторство тієї чи іншої тексту можна широко використати у багатьох сферах. Незважаючи на велику кількість досліджень, немає можливості зі стовідсотковою гарантією визначити авторство навіть художніх текстів. Однією з проблем визначення авторства художніх текстів є складність вибору параметрів тексту, які будуть визначати авторський стиль достатньою мірою

Для визначення справжнього автора тексту часто доводиться звертатися до експертів, які можуть ідентифікувати автора невідомого тексту або визначити належність твору іншому автору за допомогою характерних мовних особливостей та різних стилістичних прийомів. Експертний аналіз тексту займає багато часу і є дуже трудомістким. У зв'язку з цим велику перспективу мають формальні методи різної атрибуції текстів для автоматизації процесу аналізу.

Наразі для атрибуції текстів застосовуються підходи з теорії розпізнавання образів, математичної статистики та теорії ймовірностей, алгоритми нейронних мереж та кластерного аналізу та багато інших. Проте, всі такі методи не мають достатньої ефективності та не можуть працювати з текстами різних мов та

напрямів, а також не працюють зі стилістичними особливостями автора в достатній мірі.

Зв'язок роботи з науковими програмами, темами, планами. Підтверджено при розробці програмних засобів у рамках науково дослідної роботи кафедри «Комп'ютерні інформаційні технології» Українського державного університету науки і технологій, а саме є частиною науково-дослідних робіт «Інструментальна підтримка систем обробки природньомовних документів» (2022 р. № держреєстрації 0122U002086).

Мета та задачі дослідження. Розробити методи та програмний інструментарій для встановлення авторству природньомовних текстів та виявлення запозичень. Для досягнення мети були поставлені та вирішені задачі дисертаційного дослідження. Необхідно було розробити:

- модель представлення природньомовного тексту у вигляді правил стохастичної граматики, метод визначення авторства художніх та технічних текстів на основі цієї моделі;
- виконати експерименти для дослідження ефективності застосування стохастичної граматики на репрезентативній вибірці;
- конструктивно-продукційна модель для автоматичного будування та порівняння цих правил для встановлення подібності текстів;
- метод визначення авторству на основі декількох груп показників на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою розпізнавання образів;
- за допомогою проведення експериментів дослідити ефективність методу визначення авторству на основі декількох груп показників на репрезентативній вибірці;
- дослідити наявність зв'язку між встановленням авторства текстів та пошуком запозичень;

– метод встановлення профілю автора для подальшого його використання у визначенні авторства текстів та пошуку їх подібностей на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою багатокритеріального аналізу та методу направленого випадкового пошуку глобальних екстремумів (генетичного алгоритму) для зменшення числа показників на перелік найбільш значущих для кожного з авторів;

– провести експерименти з дослідження ефективності представленої моделі з використанням профілю автора на репрезентативній вибірці для знаходження оптимального варіанту вирішення проблеми встановлення авторства природньомовних текстів.

Об’єктом дослідження є процеси встановлення авторства природньомовних художніх та технічних текстів та визначення подібностей у них.

Предметом дослідження методи, моделі та програмні засоби для виявлення авторства природньомовних художніх і технічних текстів з оцінкою їх ефективності.

Методи дослідження. Для досягнення мети дослідження були використані наступні методи:

– адаптований до задач дослідження текстів рекурентний аналіз;

– конструктивно-продукційне моделювання для розробки представлення текстів у вигляді стохастичних граматик та їх порівняння;

– багатокритеріальна оптимізація використовувалась для роботи з різними групами показників авторського стилю;

– для встановлення авторству на основі декількох груп різних показників використовувалось розпізнавання образів;

– метод направленого випадкового пошуку глобальних екстремумів (генетичний алгоритм);

– статистичного аналізу, n-грами та аналіз складності сприйняття тексту слугували для отримання ряду показників відображення авторського стилю та використовувалися для порівняння текстів.

Наукова новизна отриманих результатів. Вперше:

– розроблено модель природньомовного тексту у вигляді множини правил стохастичних граматик та метод порівняння текстів на основі порівняння цих правил, на відміну від існуючих моделей вона дозволяє враховувати синтаксичні та стилістичні особливості тексту автора;

– розроблено метод визначення авторства тексту на основі комбінованих показників складності тексту статистичного, рекурентного аналізу та конструктивно-продукційного аналізу. Це забезпечує всебічний аналіз з врахуванням значної кількість параметрів тексту;

– розроблено метод встановлення профілю автора який визначає найбільш суттєві показники аналізу текстів притаманні певному автору, що спрощує та полегшує подальші розрахунку та скорочує потрібний час та ресурси розрахунку;

– встановлено статистично значимий зв’язок результатів рішення задач виявлення запозичень та встановлення авторству текстів.

Удосконалено метод багатокритеріальної оптимізації на основі генетичного алгоритму.

Отримали подальший розвиток:

– методи конструктивно-продукційного моделювання у частині використання параметризованих конструкторів, зв’язків між конструкторами та конструювання конструкторів;

– метод розпізнавання образів за критерієм мінімальної відстані в частині формування образу текстів;

– метод рекурентного аналізу для роботи з природньомовними текстами.

Практичне значення отриманих результатів. Результати експериментів визначили значення раціонального параметру – мінімальну довжину текстових

фрагментів (у кількості слів) яку слід вважати запозиченням, він дорівнює п’яти словам. Це слід вважати рекомендацією для використання будь-яких програм виявлення запозичень.

Запропонований метод може використовуватись як для вирішення проблем пошуку запозичень, так і для встановлення вірогідного авторства тексту. Може використовуватись як додатковий інструмент для комплексної перевірки текстів на плагіат. Особливістю є можливість використання запропонованого метода як самостійно для вирішення відповідних задач, так і комплексі з іншими інструментами для покращення їх роботи та отримання більш достовірних результатів.

Запропонований підхід може використовуватися для пошуку як повних збіжностей та співпадінь у тексті, так і схожості стилів написання та побудови речення, що може вказувати на участь декількох авторів у написанні роботи.

Впроваджено у навчальний процес та використовувалося для пошуку прямих збіжностей та подібностей стилю написання дипломних робіт студентів ОКР «Бакалавр» за напрямом 6.050103 «Програмна інженерія» ДНУЗТ–2018.

Підтверджено при розробці програмних засобів у рамках науково дослідної роботи кафедри «Комп’ютерні інформаційні технології» Українського державного університету науки і технологій, а саме є частиною науково-дослідних робіт «Інструментальна підтримка систем обробки природньомовних документів» (2022 р. № держреєстрації 0122U002086).

Особистий внесок здобувача. Основні результати дисертаційної роботи опубліковано в статтях у співавторстві:

[23, 107] – методи конструктивно-продукційного моделювання структури побудови речень у частині використання параметризованих конструкторів, зв’язків між конструкторами та їх конструювання;

[18, 19, 108] – адаптація методу рекурентного аналізу для роботи з природньомовними текстами;

[21, 22, 109] – розроблено модель природньомовного тексту у вигляді множини правил стохастичних граматик та метод порівняння текстів на основі

порівняння цих правил, на відміну від існуючих моделей вона дозволяє враховувати синтаксичні та стилістичні особливості тексту автора;

[110] – встановлено статистично значимий зв’язок результатів рішення задач виявлення запозичень та встановлення авторству текстів;

[111] – метод визначення авторства тексту на основі комбінованих показників складності тексту статистичного, рекурентного аналізу та конструктивно-продукційного аналізу;

[112] – дослідження інформативності різних одиниць тексту у задачі встановлення авторству природньомовних текстів;

[20, 113] – метод встановлення профілю автора який визначає найбільш суттєві показники аналізу текстів притаманні певному автору, що спрощує та полегшує подальші розрахунку та скорочує потрібний час та ресурси розрахунку. Представлено розроблений метод багатокритеріальної оптимізації на основі генетичного алгоритму.

Публікації. Результати дисертаційної роботи опубліковано в тринадцяти наукових працях.

У фахових та рекомендованих Міністерством освіти і науки (МОН) України для публікації результатів дисертацій – 4.

Матеріали міжнародних конференцій, що індексуються МНМБ Scopus – 3.

У тезах доповідей міжнародних та всеукраїнських конференцій – 6.

Структура та обсяг дисертації. Дисертаційна робота складається із вступу, 4 розділів, висновків, списку використаних джерел і додатків. Загальний обсяг дисертації становить 130 сторінки, в тому числі 110 сторінки основної частини, 15 рисунків, 22 таблиць, 1 додаток на 1 сторінці та список використаних джерел із 148 найменувань на 14 сторінках.

РОЗДІЛ 1
АНАЛІЗ ПІДХОДІВ ДО ВСТАНОВЛЕННЯ АВТОРСТВА ПРИРОДНЬОМОВНИХ ТЕКСТІВ

1.1 Інтелектуальний аналіз тексту

Науковий інтерес до автоматичної обробки текстів та інтелектуального аналізу текстів виник приблизно шістдесят років тому. Особливе місце в цій сфері займають проблеми виявлення авторства [92, 54, 124], плагіату [144], оцінки загальної тональності [12] та якості тексту. На цей час залишається багато невизначеного у цій проблематиці.

Проблема встановлення авторства текстів виникає у юридичній площині. Питання авторства має велике значення для усіх сфер де існує поняття права власності на об'єкт, де роль авторства є дуже істотною. Це стосується художніх творів, наукових та навчальних матеріалів та багатьох інших робіт. Складність питання полягає у тому, що для перевірки текстів на плагіат або виявлення запозичення потрібно мати відповідну базу матеріалів для порівняння [80]. Задача ускладняється багатомовністю джерел. Частково цю задачу можна вирішити без застосування матеріалів для порівняння.

Тема визначення авторства природньомовного тексту є досить актуальною. Методика, що дозволяє з достатньою достовірністю авторство тієї чи іншої тексту можна широко використати у багатьох сферах. Незважаючи на велику кількість досліджень та різноманітність підходів [58, 59, 100, 126, 49], немає можливості зі стовідсотковою гарантією визначити авторство навіть художніх текстів [68].

Однією з проблем методики визначення авторства художніх текстів є складність вибору параметрів тексту, які будуть визначати авторський стиль достатньою мірою [62, 67, 68].

Інтелектуальний аналіз тексту (ІАТ), також відомий як інтелектуальний аналіз текстових даних [1, 43] або виявлення знань з текстових баз даних [56], зазвичай відноситься до процесу вилучення цікавих і нетривіальних шаблонів або

знань з неструктурованих текстових документів. Це напрям інтелектуального аналізу даних і штучного інтелекту, метою якого є отримання інформації з колекцій текстових документів, ґрунтуючись на застосуванні ефективних методів машинного навчання та обробки природної мови [29]. Найбільша складність виникає при роботі з природною мовою або з текстами без чіткої структури контенту.

Основними сферами застосування ІАТ є інформаційний пошук, виділення інформації, категоризація та обробка природньої мови [135].

Концепція інформаційного пошуку (ІП) була розроблена у зв'язку з роботою з системами баз даних протягом багатьох років. Пошук інформації – це об’єднання запитів та пошук інформації з великої кількості текстових документів.

Завдяки величезній кількості текстової інформації інформаційний пошук знайшов багато застосувань. Існує багато інформаційно-пошукових систем, наприклад онлайн системи бібліотечних каталогів, системи керування документами в режимі онлайн та системи пошуку в Інтернеті [53].

Метод вилучення інформації визначає ключові слова та зв’язки в тексті. Реалізується шляхом пошуку попередньо визначених послідовностей у тексті так званим – зіставленням шаблону. Програмне забезпечення визначає зв’язки між усіма ідентифікованими місцями, людьми і часом, щоб надати користувачеві значущу інформацію. Ця технологія дуже корисна при роботі з великими обсягами інформації. Традиційний інтелектуальний аналіз даних передбачає, що інформація, яку шукають, вже у формі реляційної бази даних [53].

Категоризація передбачає визначення основних тем документа шляхом його зіставлення до попередньо визначеного набору тем. Класифікуючи документ, комп’ютерна програма часто сприйматиме документ як «мішок слів». Він не намагається обробити фактичну інформацію, як це робить вилучення інформації. Швидше, категоризація підраховує лише слова, які з’являються, і, виходячи з підрахунку, визначає основні теми, які охоплює документ. Категоризація часто ґрунтується на глосарій, для якого попередньо визначені теми та зв'язки

ідентифікуються шляхом пошуку великих термінів, більш вузьких термінів, синонімів і споріднених термінів [28].

1.2 Особливості обробки природної мови.

Обробка природної мови (NLP) — це область досліджень і застосування, яка досліджує, як комп'ютерні технології можна використовувати для розуміння та обробки тексту природньою мовою [77]. Дослідники NLP прагнуть зібрати знання про те, як люди розуміють і використовують мову, щоб можна було розробити відповідні інструменти та методи, щоб змусити комп'ютерні системи розуміти та маніпулювати природними мовами для виконання бажаних завдань [61, 63, 15].

Основи NLP лежать у ряді дисциплін, а саме. комп'ютерні та інформаційні науки, лінгвістика, математика, електротехніка та електронна інженерія, штучний інтелект і робототехніка, психологія тощо. Застосування NLP включає низку областей досліджень, таких як машинний переклад, обробка тексту природньою мовою та резюмування, встановлення авторству текстів та виявлення плагіату, інтерфейси користувача, багатомовний та міжмовний пошук інформації, розпізнавання мовлення, штучний інтелект та експертні системи тощо [63, 16].

Наразі більшість методів обробки природньої мови та категоризації текстів можна розділити на два широкі напрямки: застосування статистичних методів і використання машинного навчання [97].

Методи, що досліджуються можна умовно поділити на дві групи – статистичні методи аналізу тексту та машинне навчання.

Статистичні аналіз тексту працює з частотою входження різних за розміром одиниць тексту для знаходження закономірностей, що будуть характеризувати сам текст та відображати особливості його побудови та слугувати основою побудови профілю автора [90].

Для статистичного аналізу тексту можна виділити наступні рівні: частота літер, послідовностей літер деякої довжини, слів, словосполучень, речень, тощо.

Кожний з рівнів статистичного аналіз будуть мати свої особливості та точність, й можуть бути використані для відображення різних аспектів тексту, що досліджується.

Машинне навчання – це техніка аналізу даних , яка вчить комп'ютери робити те, що є природним для людей і тварин: вчитися на досвіді. Алгоритми машинного навчання використовують обчислювальні методи, щоб «вивчати» інформацію безпосередньо з даних, не покладаючись на заздалегідь визначене рівняння як модель [83]. Алгоритми адоптивно покращують свою продуктивність із збільшенням кількості зразків, доступних для навчання. Метод також широко використовується для вирішення задачі ідентифікації авторства текстів [59, 56, 100].

Модель машинного навчання для роботи з текстом включає три найпоширеніші модулі: дерева рішень (DT), нейронні мережі (NN), наївні байєсовські класифікатори та машини опорних векторів (SVM).

Кожен з представлений напрямів дослідження природньомовного тексту застосовується для вирішення різних задач та має свої особливості. Нижче розглянемо безпосередньо практичні методи кожного з напрямів аналізу тексту та можливості їх застосування.

1.2 Статистичні методи аналізу тексту

1.2.1 Частотний аналіз та аналіз на основі N-грам

Проблему статистичної та частотної структури текстів, складання частотних словників мови конкретного автора або окремо взятих текстів мовознавці досліджували для багатьох різних мов (німецькою, англійською та деякими слов'янськими мовами тощо) [27, 28, 84, 95, 96, 146, 14].

Такий аналіз ґрунтується на побудові частотного словника автора за вибраним текстом шляхом обчислення частоти входження кожного з отриманих одиниць тексту [9, 69]. Досвід складання подібних словників наочно демонструє, що словесне наповнення будь-якого, досить довгого тексту має власну

статистичну структуру. В результаті чого, можна стверджувати, що у кожного автора є співвідношення часто і рідко вживаних лексем. Саме це співвідношення читач і сприймає як багатий чи бідний словник автора [27]. Статистичні моделі мови автора також використовуються для визначення власного стилю автора у різних мовах [17, 64].

Надалі, після проведення частого аналізу, виділяються визначальні ознаки кожного з текстів. Однією з таких характеристик є авторський інваріант [31]. Це числовий параметр, який дає можливість розрізняти твір за авторським стилем. Дуже часто, як показали попередні дослідження для прози, цей показник істотно впливає частота вживання службових слів (таких як прийменники чи частки).

Частотним характеристикам текстів присвячено багато робіт, де було розглянуто подібності між авторами IXX-XX століть [2]. Також були проаналізовані подібні словники для різних слов'янських мов, таких як чеська, польська, сербська та болгарська [3].

Для атрибуції текстів використовувалися різні методи, проте найвищі результати виходять шляхом використання частотних характеристик тексту [27], N-грам [84, 58, 146] та їх різних варіацій, а також частоти слів (всіх або будь-якої окремої їх категорії [57]) та частин слів [94, 122, 126].

Серед усіх перелічених методів широко використовуваних методів аналізу тексту є метод N-грам [35]. Він часто використовується у виявленні плагіату [144], категоризації текстів [35, 82], встановленні авторства текстів [50, 125] та має ефективність на рівні 80-70%.

N-грамом в алфавіті називають довільний ланцюжок довжиною N. Як ланок такого ланцюжка можна використовувати як символи, так і окремі слова. Метод полягає в підрахунку та порівнянні профілів частоти N-грамів для різних текстів. Як показали раніше наведені дослідження, використання N-грам найбільшою мірою відображає особистий стиль автора завдяки фіксуванню послідовностей лексичних конструкцій [57, 70]. Стиль тексту багато в чому визначається частотою і порядком вживання в ньому різних частин мови [4, 147], що задовольняє умов застосування методу N-грам.

Аналіз на основі N-грамів дозволяє виявити характерні поєднання слів та їх складність для конкретного твору або автора. На основі цих даних можна визначити характерний стиль мови автора. Це твердження справедливе як звичайних, так спеціалізованих текстів [35].

1.2.2 Застосування токенізації та стемінгу у вирішенні задачі атрибуції авторського стилю та встановлення авторства тексту

Для лексичного аналізу існує процес розбиття тексту на елементарні одиниці – токени. Такий процес називається токенізацією і є зазвичай початковим етапом стемінга, адже дозволяє працювати зі словом як з окремим сутністю, при цьому знаючи його контекст. Зазвичай лексичний аналіз відбувається на рівні слів. Однак іноді буває важко визначити, що мається на увазі під «словом».

У основі перетворення морфологічних форм слова в основу здійснюється за умови, що кожна з них семантично пов’язана. Є два моменти, які слід враховувати при використанні стемеру:

- передбачається, що морфологічні форми слова мають однакове базове значення і, отже, повинні відповідати одній основі;
- слова, які мають різне значення, слід зберігати окремо.

Ці два правила є достатніми, якщо результуючі основи корисні для наших програм видобутку тексту чи обробки мови.

У мовах із відносно простою морфологією вплив коріння менший, ніж у мовах із більш складною морфологією. Більшість експериментів зі стемінгом, проведених досі, стосуються англійської та інших західноєвропейських мов.

Стемінг у широкому сенсі можна визначити як практику об'єднання семантично еквівалентних варіантів слів з одним і тим же коренем шляхом видалення словотворчих і словозмінних афіксів. З технічної точки зору, це процедура, яка намагається видалити суфікси для об'єднання варіантів слів в один.

Для англійської, як і багато західноєвропейських мов, стемінг - це переважно метод видалення суфіксів. Тобто стемінг - це процедура видалення

суфіксів, які приєднуються наприкінці слів. Тож стемінг алгоритми для англійської та інших європейських мов зазвичай не враховують префікси та інфікси. Тому стемінг насамперед пов'язаний з морфологією суфіксів.

Результати стемінгу іноді дуже схожі визначення кореня слова, та його алгоритми базуються на інших принципам. Тому слово після обробки алгоритмом стемінг може відрізнятися від морфологічного кореня слова.

Алгоритми стемінгу можна класифікувати три групи: методи усічення, статистичні методи, і змішані методи. Кожна з цих груп має типові спосіб знаходження основи варіантів слів.

Одним з найбільш поширених стемерів є стемер Мартіна Портера [6, 48]. Алгоритм набув широкого поширення і став стандартним алгоритмом стемінгу для англійської мови. Цей стемер використовує методи усічення.

Оригінальна версія стемера була призначена для англійської мови але згодом Портер використовуючи основну ідею алгоритму, написав стемер для поширених індоєвропейських мов, у тому числі для деяких слов'янських мов [137]. Існують також дослідження, що описують методи стемінгу для роботи з українськими текстами [5].

Завдяки своїм особливостям метода також має високу ефективність 75-85%.

1.2.3 Синтаксичний аналіз тексту для аналізу авторського стилю

Синтаксичний аналіз різних частин тексту є досить популярним методом аналізу самої роботи автора, її семантики, спрямованості та основний ідеї твору. Проте, цей вид аналізу текстів різних напрямів стикається зі складністю автоматичного формування синтаксичних моделей [74]. Це значною мірою обумовлено складністю структури самої мови, варіативністю використовуваних словоформ і самої структури речень. Незважаючи на це, даний метод дослідження тексту несе найбільшу кількість інформації про авторський стиль: в незалежності від тематики тексту синтаксична структура мови автора буде явно відображати його власний стиль мовлення та багато інформації можна отримати саме на основі контексту [39].

Відомі різні дослідження формалізації природної мови [32, 33, 66, 87]. Один із методів роботи з природною мовою – використання граматик [120]. Наприклад, проводилися подібні дослідження для італійської [88].

На відміну від проблеми категоризації тексту, мета якої полягає в тому, щоб визначити тему або список тем для тексту на основі його змісту, синтаксичний аналіз тексту абстрагується від конкретної області і намагається зрозуміти незалежні від змісту риси тексту [76, 103, 128], які є «лінгвістичними виразами» окремих авторів [46].

Ідея використання інформації про частини мови не нова і успішно застосовувалася в низці завдань класифікації стилів, де оброблялися, зокрема, тексти англійською мовою [60, 62, 67] та іншими [89]. Як правило, на основі частин мови витягувалися їх послідовності, що повторюються. Розуміння структури мови такою може використовуватись для покращення розуміння змісту тексту [140].

Описаний підхід добре працює для англійських текстів, оскільки структура англійської досить формальна і порядок слів у реченні чітко закріплений за певною частиною промови. Крім того, самі слова мають не так багато різних словоформ і при додаванні або видаленні префікса або суфікса переходять в розряд іншої частини мови. Проте при використанні такого методу для інших мов можуть виникнути труднощі, тож робота з будь-якою мовою потребує врахування її особливостей, що робить неможливим створення універсального інструменту та вимагає кропіткого налаштування під кожен окремий випадок, маючи ефективність 75-90%.

1.3 Застосування машинного навчання у визначенні авторства текстів

1.3.1 Дерева рішень як інструмент класифікації

Дерева рішень є парадигмою машинного навчання, спеціально розробленого для підтримки описової класифікації та пояснення чому класифікація відбулася саме так.

Ця техніка на основі дерева, в якій будь-який шлях, починаючи з кореня, описується послідовністю розділення даних до тих пір, поки не буде досягнуто логічний результат у листовому вузлі [145]. Це ієрархічна ілюстрація зв'язків знань, які містять вузли та зв'язки. Коли відношення використовуються для класифікації, вузли представляють цілі [72].

Дерева рішень є одним із найпотужніших методів та використовується в різних областях, таких як машинне навчання, обробка зображень та ідентифікація шаблонів. Це послідовна модель, яка об'єднує серію базових тестів ефективно та узгоджено, де числова характеристика порівнюється з пороговим значенням у кожному тесті [41]. Концептуальні правила набагато легше побудувати, ніж числові вагові коефіцієнти в нейронній мережі зв'язків між вузлами [30, 52]. В основному використовується для групування, але крім того, зазвичай використовуваною моделлю класифікації в Data Mining [47]. Вузли та гілки складаються з кожного дерева. Кожен вузол представляє ознаки в категорії, що підлягає класифікації, і кожна підмножина визначає значення, яке може прийняти вузол [42]. Завдяки своєму простому аналізу та їх точності на багатьох формах даних, дерева рішень знайшли багато полів реалізації [91], у тому числі при роботі з природньомовними текстами [36].

У сфері обробки природньої мови, згідно існуючим дослідженням ефективність методу сягала 75-85% при роботі з різними мовами та напрямками текстів.

1.3.2 Застосування нейронних мереж у задачі встановлення авторства природньомовних текстів

Нейронні мережі основуються на моделі людського мозку в якості метафори [99]; вони складаються з великої кількості взаємодіючих простих арифметичних процесорів.

Як і у випадку з біологічними мережами, окремі вузли в штучних нейронних мережах називаються нейронами. Ці нейрони є обчислювальними одиницями, які отримують вхідні дані від інших нейронів, виконують обчислення

на цих вхідних даних і передають їх іншим нейронам. На обчислення в нейроні впливають вагові коефіцієнти вхідних з’єднань цього нейрона, оскільки вхідні дані нейрона масштабуються за ваговими коефіцієнтами. Ці вагові коефіцієнти можна розглядати як аналог міцності синоптичного зв’язку. Відповідним чином змінюючи ці вагові коефіцієнти, можна вивчити загальну обчислювальну функцію штучної нейронної мережі, що є аналогом вивчення синоптичної сили в біологічних нейронних мережах [26]. Ідея полягає в тому, щоб поступово змінювати вагові коефіцієнти щоразу, коли поточний набір вагових коефіцієнтів робить неправильні прогнози [26].

Існують дослідницькі роботи, яким вдалося перетворити такі мережі на набори правил, щоб дізнатися, чого навчилася мережа [44, 102], однак багато інших робіт все ще називають такий підходом «чорної скриньки» [102, 131], через труднощі в розумінні процесу прийняття рішень мережевою мережею, що може призвести до невідомості про успішність тестування. Метод використовують для роботи з текстами на різних мовах [40, 55, 56] та при роботі зі структурою речень [55, 75, 138, 141], у тому числі українською [79]. Ефективність методу має широкий розбіг – від 74 до майже 92%.

На жаль, вони не можуть перевершити інші методи, в першу чергу опорний вектор машини, на даний момент, мабуть, найточніший з відомих методів класифікації [78].

1.3.3.Застосування наївного методу Байєсу при класифікації текстів

Наївний Байєс – це простий імовірнісний класифікатор, який оцінює набір ймовірностей шляхом обчислення частоти та розташування значення у наборі даних [134].

Наївні байєсівські класифікатори виконують аналогічну класифікацію, але без деревоподібної структури. Натомість вони покладаються на простий обчислювальний застосування теореми Байєса для виведення ймовірності того, що схема класифікації виведе найбільш ймовірну категорію з урахуванням спостережуваних даних. Їх називають «наївними», тому що вони використовують

прості і нереалістичні припущення про незалежність (наприклад, вони можуть припустити, що частина слова «я» не залежить від частоти слова «я», свідомо хибне припущення), але, тим не менш, може працювати напрочуд добре і надзвичайно швидко розвивається і тренуватися [143].

Наївні моделі Байєса – моделі популярні в додатках машинного навчання через їхню простоту, що дозволяє кожному атрибуту вносити свій внесок у остаточне рішення однаково й незалежно від інших атрибутів. Ця простота прирівнюється до обчислювальної ефективності, що робить метод привабливим та придатним для багатьох областей. Однак те саме, що робить їх популярними, також є причиною, яку наводять деякі дослідники, які вважають цей підхід слабким, однак, якщо вони використовуються у відповідних областях, вони пропонують швидке навчання, швидкий аналіз даних і прийняття рішень, а також проста інтерпретація результатів тесту [143]. Існують дослідницькі роботи [73, 139], які намагаються пом’якшити припущення про незалежність змінних шляхом введення прихованих змінних у їхніх деревоподібних або ієрархічних класифікаторах.

Удосконалення стандартного правила або його використання у співпраці з іншими методами може значно покращити результати [25]. Наприклад, NBTree [139] досліджують роботу гібриду, використовуючи правило Байєса для побудови дерева рішень. Інші дослідницькі роботи [73] модифікували свої класифікатори, щоб навчатися на позитивних і немаркованих прикладах.

Алгоритм легко і швидко передбачає клас тестового набору даних. Він також добре справляється з багато класовим прогнозуванням. Продуктивність наївного байєсовського класифікатора краще, ніж в інших простих алгоритмів, таких як логістична регресія і вимагає менше навчальних даних. Обмеженням даного алгоритму є припущення про незалежність ознак. Однак у реальних завданнях цілком незалежні ознаки трапляються вкрай рідка.

Метод має середню ефективність у порівнянні з іншими, всього 70-80%.

1.3.4 Метод опорних векторів у задачі атрибуції природньомовних текстів та встановленні їх авторства

Машини опорних векторів [132] мають сильні теоретичні основи та чудові емпіричні успіхи. Вони застосовувалися для таких завдань, як розпізнавання рукописних цифр, розпізнавання об’єктів і класифікація тексту.

Хоча здатність до навчання та обчислювальна складність навчання на опорних векторних машинах може не залежати від розмірності простору однак, зменшення обчислювальної складності є важливою проблемою для ефективної обробки великої кількості термінів у практичних застосуваннях класифікації тексту [65].

Машина опорних векторів є технікою класифікації яка прагне знайти гіперплощину, що розбиває дані за їх міткою класу і в той же час уникати надмірної фільтрації даних [34]. Вивчення гіперплощини в лінійній здійснюється шляхом перетворення задачі за допомогою лінійної алгебри [121]. І це робить двійкову класифікацію, засновану на розділенні гіперплощини на повторно відображеному просторі екземплярів.

Цей метод є одним із найвідоміших методів оптимізації очікуваного рішення [93, 133]. Його надзвичайна здатність до узагальнення разом із оптимальним рішенням і розрізняльною здатністю привернула увагу інтелектуального аналізу даних, розпізнавання образів і машинного навчання в останні роки. Було показано, що SVM перевершують інші методи навчання під наглядом [97]. Завдяки хорошій теоретичній основі та хорошій здатності до узагальнення SVM стали одним із найбільш використовуваних методів класифікації. Ефективність застосування становить 77-90%.

Висновки по першому розділу

Вперше закладено основу для створення інструменту з аналізу побудови україномовних текстів за допомогою систематизації існуючих методів та

розробок, що працюють для інших мов. Виявлено можливості роботи з тестами наступних методів:

статистичних:

- частотний аналіз [26, 30, 31, 43, 53];
- стеммінг [63];
- синтаксичний аналіз [66, 67, 69, 107, 109, 136, 144];

методів машинного навчання:

- дерева рішень [78, 84, 93, 95, 96];
- нейронні мережі [81, 97, 99, 102, 146];
- наївні байесовські класифікатори [121, 122, 127, 130, 131];
- машини опорних векторів [58, 132, 133, 135, 137].

Виконаний аналіз дозволив виділити найбільш значні методи роботи з природньомовними текстами. Закладено теоретичні основи для створення інструменту для роботи з україномовними текстами.

Було проведено аналіз найефективніших методів, що найчастіше використовуються для аналізу та атрибуції тексту та визначення авторства.

Однак, незважаючи на необхідність створення окремого інструменту для роботи саме з українською мовою, методи є ефективними. Результат визначення авторства знаходиться в діапазоні від 74% до 92% правильно встановлених випадків. Дані результати варіювалися в залежності від використовуваного методу, мови і стилю аналізованого тексту.

За результатами проведеного дослідження можна стверджувати, що кожен з методів не є універсальним та має свої недоліки. Крім того, робота з різними мовами вимагає урахування їх особливостей безпосередньо у побудові слів та речень. Це вимагає модифікацію існуючих методів та потребує розробки окремого підходу для роботи з урахуванням власних особливостей україномовних текстів. Серед великої кількості розглянутих інструментів жодний не має стовідсоткової ефективності у вирішенні задачі встановлення авторству тексту.

За матеріалами розділу опубліковано роботи [107, 108, 109, 110, 111].

РОЗДІЛ 2
МЕТОДИ ТА МОДЕЛІ ВСТАНОВЛЕННЯ АВТОРСТВА ПРИРОДНЬОМОВНИХ ТЕКСТІВ

2.1 Адаптований для роботи з украънськими природньомовними текстами рекурентний аналіз

Рекурентний аналіз використовується для дослідження часових рядів та роботи зі складними системами [85, 86]. Він був модифікований для аналізу текстів.

За основу був узятий аналіз рекурентних діаграм (recurrence quantification analysis, RQA), в якому для аналізу використовують щільність рекурентних точок [148].

Модифікований метод для роботи з природньомовними текстами полягає у наступному:

– розраховується частота входження текстової одиниці (літери або їх послідовності) у тексті;

– отримується часовий ряд, замінюючи кожний символ обраного тексту на його частоту. Умовний час – перехід від одного символа до іншого;

– визначається фазовий простір [129], як візуалізація переходів від стану до стану (від символа до символа);

– розраховується рекурентна діаграма на основі фазового простору через відображення повторюваних станів у різні моменти часу;

– обчислюються та інтерпретуються загально вживані показники рекурентного аналізу щодо аналізу тексту.

Показник рекурентності (recurrence rate, *RR*) визначає щільність рекурентних точок на досліджуваній діаграмі. Це значення приблизно відображає загальну кількість повторень кожного з статистично близьких символів

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}^{m,\varepsilon}, \qquad (2.1)$$

де N – кількість розглянутих станів, $R_{i,j}$ – i,j-та точка рекурентної діаграми, ε – радіус околиці точки в момент часу i, m – розмірність фазового простору.

Показник детермінізму (determinism, *DET*) розглядає частотний розподіл довжин l діагональних ліній в діаграмі $P^{\varepsilon}(l)$, де N – абсолютна кількість таких ліній. Значення *DET* визначає частоту повторень всіх сполучень статистично близьких символів будь-якої довжини:

$$DET = \frac{\sum_{l=l_{\min}}^{N} lP^{\varepsilon}(l)}{\sum_{i,j}^{N} \mathrm{R}_{\mathrm{i,j}}^{m,\varepsilon}}. \qquad (2.2)$$

Середня довжина діагональних ліній L визначає середню довжину повторюваних статистично близьких символів.

$$L = \frac{\sum_{l=l_{\min}}^{N} lP^{\varepsilon}(l)}{\sum_{l=l_{\min}}^{N} P^{\varepsilon}(l)}. \qquad (2.3)$$

Показник дивергенції (divergence, *DIV*) є величиною, зворотною максимальній довжині діагональних структур.

$$DIV = \frac{1}{\max(\{l_i; i = 1...N_l\})}. \qquad (2.4)$$

Ентропія (entropy, *ENTR*) є показником частотного розподілу діагональних ліній, для текстів – частотного розподілу повторюваних поєднань статистично близьких символів.

$$ENTR = -\sum_{l=l_{\min}}^{N} p(l)\ln(p), \qquad (2.5)$$

де

$$p(l) = \frac{P^{\varepsilon}(l)}{\sum_{l=l_{\min}}^{N} P^{\varepsilon}(l)}. \qquad (2.6)$$

Показник завмирання (laminarity, *LAM*) демонструє частотний розподіл довжин v горизонтальних ліній в діаграмі $P^{\varepsilon}(v)$, де N – абсолютна кількість таких ліній. Показник *LAM* приблизно визначає повторення статистично близьких символів.

$$LAM = \frac{\sum_{v=v_{\min}}^{N} vP^{\varepsilon}(v)}{\sum_{i,j}^{N} \mathrm{R}_{\mathrm{i,j}}^{m,\varepsilon}}. \quad (2.7)$$

Показник затримки (trapping time, *TT*) відображає середню довжину горизонтальних структур. Показник *TT* визначає середню довжину поєднань статистично близьких символів.

$$TT = \frac{\sum_{v=v_{\min}}^{N} vP^{\varepsilon}(v)}{\sum_{v=v_{\min}}^{N} P^{\varepsilon}(v)}. \quad (2.8)$$

Показники (1)..(8) відображають структуру рекурентної діаграми.

2.3 Моделювання тексту засобами формальних стохастичних граматик

Генеративні (або трансформаційні) граматики були розроблені лінгвістами [45, 37, 38] з метою вивчення структури природних і формальних мов. Граматики описують синтаксичну структуру речення (або рядка) формальною мовою. Граматика містить набір правил перетворення, які рекурсивно описують мову.

Граматика є неоднозначною граматикою, якщо існує кілька шляхів для одного рядка. Призначення ймовірностей рядкам і їх різним шляхам синтаксичного аналізу можна використовувати як спосіб створення багатозначності різних правил. Стохастична граматика пов’язує ймовірності його спрацювання з кожним правилом.

Стохастична граматика використовується для створення правил, які описують структуру речень у тексті. Для кожного правила визначається ймовірність його застосування в конкретній роботі. Імовірність виведення всього речення визначається як добуток ймовірностей послідовностей частин мови, які використовуються в ньому. Отримані правила створюють мову, специфічну для досліджуваних і структурно схожих творів певного автора.

Для опису основної структури тексту використовувалися частини мови як характеристика слова. Таким чином, кожне слово в реченні замінюється частиною мови, якою воно є.

Кожне зі слів у тексті було проаналізовано на предмет схожості з наявними в українській мові частинами мови. Для службових частин мови: прийменників, займенників, сполучників і вставних слів використовувався їх перелік у всіх можливих формах, а дієслова, іменники, прислівники, прикметники, дієприкметники та частки визначалися зіставленням із переліком закінчень слів.

Коли в одному зі списків було знайдено відповідне слово або його закінчення, слово автоматично замінювалося на відповідну частину мови. Якщо автоматично визначити відповідь не вдавалося, користувачеві пропонувалося потім включити все слово або його закінчення (дані вводилися вручну користувачем) у вже існуючий список.

Для позначення слів у тексті українською мовою використовувалися такі теги: дієслово (v), іменник (n), займенник (prn), прикметник (adj), сполучник (cnj), прислівник (adv), прийменник (prp), частка (prtcpl), вставне слово (intrj), герундій (ger).

Для кожної частини мови розраховується ймовірність її появи в певному місці речення в даному тексті. Імовірність появи певної частини мови в досліджуваній послідовності дозволить точніше вловити індивідуальний стиль письма кожного з досліджуваних авторів. Після отримання тексту у вигляді послідовностей частин мови, встановлених у реченнях з імовірністю їх появи в певному місці, формуються правила.

Для цього всі речення, що починаються з однієї частини мови, групуються, перше слово викидається, а для наступного слова повторюється процедура обчислення ймовірності.

Після того, як речення знову групуються відповідно до частин мови на початку, перше слово знову відкидається, а ймовірність для наступного елемента обчислюється і так далі. Імовірність розраховується як кількість випадків у тексті, поділена на їх загальну кількість.

Таким чином, правила підстановки для деякого тексту T мають початковий нетермінал, потім термінали, що відповідають кожному слову в реченні, і ймовірність застосування відповідного правила при розборі тексту і мають вигляд:

$$\sigma \xrightarrow{p_{1j}} b_{1j} A_{1,j}, \quad A_{i,j} \xrightarrow{p_{i+1,k}} b_{i+1,k} A_{i+1,k}, \quad j = 1 \ldots J_i, \; k = 1 \ldots K_i$$

де σ – початковий нетермінал; b_{ij} – термінали, відповідні -му слову в реченні (і відповідні і-тому правилу при розборі даного твору або і-тому рівню правила), $A_{i,j}$ – j-тий нетермінал у правилі і-го рівня, p_{ik} – ймовірність використання відповідного правила, J_i, K_t – кількість різних нетерміналів у правій частині правил і-го рівня та j-го рівня, відповідно.

Рівень відповідає порядковому номеру слова у реченні.

Допускається кілька альтернативних правил з нетерміналом у лівій частині правила, але при цьому термінали у правій частині таких правил різні, що забезпечує детермінований розбір.

Таким чином, текст представляється у вигляді набору правил, що описують його структурні особливості за допомогою описаних вище правил. Символ ε означає пусто (кінець правила).

2.4 Формування профілю автора

2.4.1 Статистичний аналіз тексту

Метод аналізу текстів із застосуванням N-грам є порівняно новим методом і найчастіше використовується для пошуку плагіату у різних текстових джерелах. Даний метод також показує добрі результати у визначенні авторства текстів.

При аналізі тексту на основі N-грам встановлюється довжина послідовності символів, частота яких використовується для визначення авторства. Від довжини символів залежить точність методу. Для автора будується профіль частоти використання різних N-грамів, що дозволяє відобразити особливості його мови.

Згідно з проведеними раніше дослідженнями найбільшу достовірність показав аналіз текстів із використанням 4-грамів.

У роботі використовується саме ця довжина послідовності символів, а розбиття слів виконується внахлест - кожен наступний 4-грам відрізняється від попереднього лише одним символом. Всі розділові знаки, регістр літери та різні елементи форматування при формуванні 4-грам не враховуються.

Наприклад, для фрази "Він йшов до мене" при її розбитті на 4-грами отримаємо наступний список: Вінй, інйш, нйшо, йшов, шовд, овдо, вдом, омен, мене.

Цей підхід дозволяє аналізувати як використовувані у мові автора слова, а й у певною мірою відображає їх послідовність.

Було створено два різні словники. Першим словником (based on VESUM dictionary) став словник загальнодоступний the Large Electronic Dictionary of Ukrainian (VESUM) [51].

На його основі було збудовано комплексний словник, що містить унікальні основи слів, їх закінчення та префікси. Для зменшення його розміру було проведено попередню вибірку унікальних списків закінчень та присвоєння основі слова лише індексу з нього. Ведення списку чергування голосних у словах також підтримується.

Для створення списків префіксів для основ проведено аналіз сформованого словника на наявність основ, які відрізняються лише наявністю префікса простим перебором. В результаті початковий словник основ зменшився – всім ключовим основам присвоєно відповідний індекс зі списку приставок, а зайві основи з приставками видалено.

Перевагою отриманого словника є підтримка обліку всіх словоформ для основ, кожної з яких буде присвоєно унікальний індекс. Таким чином, усі відмінки, різні форми слів, а також отримані додаванням приставки слова безпомилково будуть вести до єдиної основи.

Для другого словника використовувалися тексти різних стилів: художня література, публіцистика, наукові тексти, офіційно-ділові документи та тексти з

використанням розмовної лексики. Кожен із словників складається з трьох частин.

Перша включають список основ різних слів, друга - закінчень, третя - префіксів. Кожній основі відповідає список усіх можливих закінчень та приставок для цього слова. Відображається також чергування голосних у слові у процесі зміни форми.

Список основ у файлі словника складено в алфавітному порядку, за кожною основою слід три числові елементи. Перше число – номер рядка з файлу із закінченнями, у якому перераховані всі, зустрічаються з цією основою. Якщо під час використання даного закінчення основу відбувається чергування голосних, перед ним є символ «+». Друге місце займають номер літери в основі з її початку, яка бере участь у чергуванні, та голосна для заміни. Третій елемент – номер рядка із файлу з префіксами, характерними для цієї основи. Якщо якийсь із елементів списку відсутній, його замінює «-».

Приклад основ:

• «важливий 0 – 20» (є закінчення, немає чергування, є префікс);

• «абсолютніст 41 8о -» (є закінчення, 8-а буква змінюється на о, немає префікса);

• «полудні - - 7» (немає закінчення, ні чергування, є префікс).

Приклад закінчень для наведених вище основ (1 та 2):

• «а, е, ий, їм, ними, їх, і, ій, ім, ого, ої, ому, ою, у»;

• «ь, ю+е+і+і».

Приклад префіксів для наведених вище основ (1 та 3):

• «архі, гіпер, мега, над, супер, ультра»;

• "про".

Надалі дані словника використовувалися щоб прорахувати частоту входження тієї чи іншої основи в стилі автора і скласти профіль його мови.

Покриття словників складає від 93 до 96% слів.

Також застосований адаптований до української мови стемер Портера для роботи безпосередньо з текстами різних авторів та також побудовою профілю частоти використання різних основ, характерний для кожного автора.

Алгоритм стемінгу послідовно застосовує ряд правил відсікання постфіксів, закінчень та суфіксів, без використання основ. На вхід подається слово та класи морфем для різних частин мови. Алгоритм, реалізований у цій роботі, складається з наступних кроків:

1. попередня обробка слова (приведення до нижнього регістру, заміна апострофа та «ґ»);
2. перевірка є слово інфінітивом, якщо так – алгоритм завершує роботу, інакше – перехід до наступного кроку.
3. виділення частини слова після першої голосної (RV);
4. видалення з RV морфеми (суфікс, закінчення, постфікс). Якщо видалити морфему певної частини мови не вдалося, алгоритм переходить до видалення морфеми наступної частини в такому порядку:
    a. дієслова (постфікси –ся, сь);
    b. прикметники та причастя;
    c. дієслова
    d. іменники;
5. видалення з RV "і";
6. видалення з RV словотворних морфем (послідовність з 1-3 голосних на початку, що закінчується на "-сть");
7. видалення з RV подвоєння приголосних «нн» та м'якого знака;
8. відтворення слова з частини до першої голосної включно та RV – утворення стеми.

2.4.2 Показники складності сприйняття тексту

Показники складності сприйняття тексту. Лексику прийнято вважати найкращим показником легкості сприйняття тексту. Середня довжина слів (в

буквах або символах) і речень є статистичними факторами, які часто використовують для оцінки складності тексту. Ці параметри легко піддаються кількісному вираженню і придатні для автоматичної оцінки.

Проблему визначення складності тексту для розуміння читачем допомагають вирішити цілий ряд показників. На-приклад, індекси туманності Ганнінга, Колемана-Лиау та оцінка читабельності Рейгора [11]. Вони будуються на основі підрахунку кількості речень, слів, складів, букв у тексті, також середньої кількості слів, складів, букв у реченнях, та складів і букв у словах.

Усі перелічені вище показники розраховувались для текстів англійської мови вузького призначення та для певної аудиторії читачів [11]. Тому вони не зовсім відповідають цілі дослідження, однак початкові кількісні показники мають певну інформативність.

З використанням цього методу аналізу тексту заголовки, підзаголовки і формули найчастіше ігноруються, оскільки не є повноцінними реченнями.

Ці дані також несуть у собі певну інформацію про авторський стиль листа. Однак подібні показники враховують складність тексту, але не відображають його зміст і порядок слів, тому дані показники не мають достатньої ефективності для аналізу авторського стилю самостійно, але можуть використовуватися спільно з іншими показниками.

Ступінь складності текстів може давати відповідну характеристику автора.

2.4.3 Оптимізація розпізнавання образів засобами генетичного алгоритму

Генетичні алгоритми використовуються для розв'язання задачі оптимізації значення багатопараметричних функцій. Усі представлені завдання формуються як функції, які залежать від деякої кількості параметрів, глобальний максимум або мінімум яких відповідатиме вирішенню задачі.

Ідея генетичного алгоритму – організація еволюційного процесу отримання кінцевого оптимального рішення [7, 98, 114, 142]. У ньому зберігається біологічна термінологія. Таким чином, хромосома це вектор, кожна позиція якого називається геном. Кожен такий вектор (особина) характеризується певною

функцією здоров'я (функцією пристосованості). Ця функція визначає якість представленого рішення. Завдання оптимізації може розглядатися як завдання пошуку особи з найкращою функцією здоров'я. Пошук ґрунтується на механізмах спадковості, мінливості, відбору та реалізуються за допомогою різних генетичних операцій. Кросовер – операція, коли він дві хромосоми обмінюються своїми частинами. Мутація – випадкова зміна однієї чи кількох позицій у хромосомі.

Працюючи з генетичним алгоритмом початкова популяція генерується зазвичай випадково. Єдиний критерій – достатня різноманітність особин, щоб уникнути попадання популяцій до найближчого локального екстремуму.

Після генерації першого покоління генетичний алгоритм імітує еволюційний процес як процес схрещування і мутацій, що повторюється, ймовірність участі особини в схрещуванні прямо пропорційна її здоров'ю. Результатом ставатиме нова популяція, а стара гине, таким чином, функція здоров'я всіх особин від покоління до покоління в середньому покращується. Згодом процес повторюється до того часу, поки функція здоров'я припинить поліпшуватися. Як результат вибирається особина з найкращим показником функції здоров'я з останнього покоління особин.

У цьому роботі генетичний алгоритм матиме такі характеристики:

– фіксований розмір популяції;

– фіксована розрядність генів;

– пропорційний відбір;

– особини для схрещування вибираються серед найкращих представників популяції;

– схрещування;

– нащадки посідають місце попередньої популяції;

– до кожної популяції додається фіксована кількість випадково згенерованих особин для уникнення виродження популяції.

Функція здоров'я (пристосованості) визначає кількість правильно встановленого авторства для текстів навчальної вибірки.

Початкова популяція згенерується випадковим чином. Під час імітації еволюційного процесу, відбір особин наступного покоління здійснювався за такими пропорціями: 34% батьківський особин з найкращими показниками функції здоров'я схрещувалися між собою, 60% залишилися батьківських особин мутували випадковим чином, 6% особин нащадків генерувалися випадковим чином для. В експерименті розмір вибірки фіксовано і становить 100 особин у кожному поколінні.

2.4.4 Визначення авторства тексту методом порівняння з еталоном теорії розпізнавання образів

Для визначення авторства тексту використовується теорія розпізнавання образів [123], а точніше метод розпізнавання за найменшою відстанню до еталону [108].

Нехай є M класів образів $\omega_1, \omega_2, \ldots, \omega_M$, кожен з яких асоціюється з конкретним автором та образ X_l тексту, авторство якого необхідно встановити. Відомо, що цей текст належить одному з авторів.

Загальний образ тексту складається з послідовного поєднання показників різних класів у єдиний вектор, що репрезентує особливості мовленнєвого стилю автора.

Наприклад, нижче представлений зіставлений вектор (X_{4j}), що поєднує у собі наступні класи показників: частота літер (X_{1j}), довжина слів (X_{2j}), показники рекурентного аналізу (X_{3j}). Елементи векторів x_{1jk} – k-ий показник рекурентного аналізу; x_{2jk} – частота k-ї літери у тексті; x_{3jk} – кількість слів довжиною k-літер.

Згідно тексту «Заповіту» отримані значення векторів X_{1j}, X_{2j}, X_{3j}, X_{4j}.

$X_{1,1} = [0.02 \quad 0 \quad 0.13 \quad 2.38 \quad 0.77 \quad 0 \quad 2]$; $X_{2,1} = [0.06 \quad 0.02 \quad 0.06 \quad 0.02 \quad \ldots \quad 0.02]$;

$X_{3,1} = [14 \quad 11 \quad 2 \quad 18 \quad 11 \quad 11 \quad 7 \quad 4 \quad 4 \quad 1]$;

$X_{4,1} = [0.02 \quad 0 \quad 0.13 \quad 2.38 \quad 0.77 \quad 0 \quad 2 \quad 0.06 \quad 0.02 \ldots \quad 0.02 \quad 14 \quad 11 \quad 2 \quad 18 \quad 11 \quad 11 \quad 7 \quad 4 \quad 4 \quad 1]$.

Визначено зразки образів кожного класу $Z_1, Z_2, \dots, Z_M$. Еталоном класу є вектор, що містить середні значення кожного з показників текстів автора в навчальній вибірці.

У попередніх дослідженнях вважалося, що образ X_l належить класу ω_i (текст – i-му автору) якщо $\rho(X_l, Z_i) < \rho(X_l, Z_j)$ при $\forall j \neq i$, где $\rho(X, Z)$ – відстань між образами X та в Евклідовому просторі.

Різні показники можуть мати різні одиниці виміру та шкали. Для вирішення цієї проблеми використовується мінімаксна нормалізація кожного показника у векторах X та Z.

Образ тексту містить перелік показників, інформативність кожного завдання розпізнавання образів різна. У зв'язку з цим для підвищення ефективності розпізнавання вирішено використовувати вагові коефіцієнти показників. При цьому вирішальна функція має вигляд:

$$d_{lm} = \sum_{i=1}^{N} w_i (x_{il} - z_{im})^2, \tag{2.9}$$

де i – номер показника у векторі;

l – номер досліджуваного тексту, l = 1..L;

m – номер еталона автора;

w_i – ваговий коефіцієнт i-того показника;

N – кількість показників у векторі (образі тексту);

x_{il}, z_{ik} – компоненти векторів X_l и Z_k.

Вважатимемо, що образX_l належить классу ω_i (текст – i-му автору) якщо $d(X_l, Z_i) < d(X_l, Z_j)$ при $\forall j \neq i$.

Завдання полягає в тому, щоб знайти такі вагові коефіцієнти показників w_i, щоб точність розпізнавання була максимальною. Для вирішення цього завдання застосовувався генетичний алгоритм.

2.4.5 Розрахунок граничних значень довірчого інтервалу за допомогою критерію Стьюдента

Для більш достовірних результатів виконувався розрахунок довірчих інтервалів кожного з авторів вибірки. Застосовувався критерій Стьюдента [32].

Для розрахунку довірчого інтервалу в навчальній вибірці для кожного з авторів було прораховано схожість представлених текстів один до одного. Дані про схожість текстів усередині навчальної вибірки кожного з авторів ділився на частини з однаковою кількістю складових.

Для розрахунку довірчого інтервалу використовувалась формула:

$$t_{2,\beta}\sqrt{\frac{1}{6}\sum_{k=1}^{3}(\zeta_k-\theta_s)^2},$$

де $t_{2,\beta}$ – коефіцієнт Стьюдента, β – рівень довіри, ζ_k – середнє значення k-тої частини вибірки, θ_S – середнє значення по усій вибірці.

2.5 Конструктивно-продукційне моделювання авторських текстів

Конструктивно-продукційне моделювання ефективно використовувалось для рішення цілої низки практичних задач, таких як: дослідження розподілу відновної енергії [101], моделювання природної мови [116, 117], побудова розкладу занять [24], онтологічне забезпечення [106], дослідження процесів розробки та налагодження комп'ютерних програм [118, 119], моделювання геометричних фракталів [105], моделювання грозового фронту [104] та інших.

2.5.1 Узагальнений конструктор

Розвинемо конструктивно-продукційний підхід для вирішення задачі встановлення авторства технічного тексту. Узагальненим конструктором C_G називається трійка [115]

$$C_G = \langle M, \Sigma, \Lambda \rangle, \tag{2.10}$$

де M – неоднорідний носій структури, що розширюється в процесі конструювання; Σ – сигнатура операцій та відношень, що складається з операцій

зв'язування, підстановки та виведення, операцій над атрибутами та відношення підстановки; Λ – інформаційне забезпечення конструювання (ІЗК).

Згідно з Λ [115] формою ${}_{w}l$ з атрибутом w називається набір терміналів та нетерміналів, що об'єднуються операціями зв'язування. У розроблених конструкторах використовується єдина операція (та відношення) зв’язування – конкатенації. Конструкцією називається форма, яка містить тільки термінали. Формування конструкцій відбувається шляхом виведення з початкового нетерміналу, виконання операцій підстановки та операцій над атрибутами та узагальненими операціями часткового та повного виводу.

Операція часткового виведення ($|\Rightarrow\in \Sigma_{\mathrm{P}}$) складається з вибору відповідного правила підстановки з їх множини, виконання цієї підстановки та виконання операцій над атрибутами, що відповідають обраному правилу в певній послідовності.

Операція повного виведення (або просто виведення, $||\Rightarrow\in \Sigma_{\mathrm{P}}$) полягає у послідовному виконанні операції часткового виведення, починаючи з початкового нетерміналу та закінчуючи конструкцією.

Для формування конструкцій необхідно виконувати низку уточнюючих перетворень конструкторів:

- спеціалізація – визначає предметну область: семантичну природу носія, скінчену множину операцій та їх семантику, атрибутику операцій, порядок їх виконання та обмеження на правила підстановки;
- інтерпретація – полягає у зв'язуванні операцій сигнатури з алгоритмами виконання деякого алгоритмічного конструктора;
- конкретизація – розширення аксіоматики множиною правил продукцій, завдання конкретних множин нетермінальних та термінальних символів з їх атрибутами та, за необхідності, значень атрибутів;
- реалізація – формування конструкції з елементів носія конструктора шляхом виконання алгоритмів, пов'язаних з операціями сигнатури.

2.5.2 Конструктор-перетворювач природньомовного тексту у тегований текст

Метою конструювання є перетворення технічного тексту на тегований. Для кожного слова у реченні визначається його атрибутика у складі: частина мови (pos), число (num) та рід (gen). Розглянемо спеціалізацію конструктора:

$$C = \langle M, \Sigma, \Lambda \rangle \ _S \mapsto C_{\text{P}} = \langle M_{\text{P}}, \Sigma_{\text{P}}, \Lambda_{\text{P}} \rangle, \quad (2.11)$$

де M_{P} – носій, що включає термінальний та нетермінальний алфавіт, початковий та тегований тексти, а також множину правил продукції Ψ, окремі правила ψ_i:<s_i, g_i>, де i – номер правила, s_i – послідовність відношень підстановки, g_i – послідовність операцій над атрибутами, Σ_{P} – операції та відносини на елементах M_{P}; ІЗК $\Lambda_{\text{P}} \supset \Lambda$.

ІЗК Λ_{P} включає представлені далі положення.

Сігнатура Σ_{P} містить сигнатуру конкретних операцій зв’язування та операцій над атрибутами.

До терміналів T належать символи та слова української мови, позначені як * – літери, що можуть використовуються для формування слів, $\underline{*}$ – пробіл, $\overline{*}$ – символи кінця речення, ⊥ – символ кінця тексту, $W_{i,j}$, - j-те слово у i-му реченні. У перше слово кожного речення буде додатково зберігатися інформація стосовно його довжини l, а у найперше слово тексту – кількість слів у найдовшому реченні max та загальна кількість речень у тексті S. Нетермінали $N = \{\sigma, \eta, \varepsilon\}$ – допоміжні елементи, де ε – символ «пусто».

Надалі представлені операція над атрибутами.

Операція $\odot\ (word,\ ends, pos \downarrow word) \in \Sigma_{\text{P}}$ – визначення для слова *word* його частини мови *pos*, що приймає одне з наступних значень: дієслово (v), іменник (n), числівник (*nume*), займенник (*pron*), прикметник (adj), спілка (*conj*), прислівник (adv), прийменник (*prep*), дієприкметник (v_adv), вигук та частка (*frac*), дієприслівник (v_adj).

Операція $\circledast$ ($word, ends, num \downarrow word$) $\in \Sigma_{\text{P}}$ – визначення для word його числа num, що може бути однина ($sing$) або множина ($plur$).

Операція $\circledcirc(word, ends, gen\downarrow word) \in \Sigma_{\text{P}}$ – визначення для *word* його роду *gen*, що приймає одне з наступних значень: жіночій (*f*), чоловічий (*m*) та середній (*n*).

Кожна з цих операцій виконує порівняння *word* з усіма елементами $ends$ – відповідних списків закінчень [10], при співпадінні з певним закінченням формується результат та параметрам *pos, num* та *gen* будуть надані відповідні значення.

Операція = $(a, b) \in \Sigma_{\text{P}}$ – присвоєння значення b змінній a.

Операція $+(c, a, b) \in \Sigma_{\text{P}}$ – додавання $c = a + b$.

Операція порівняння <>*(a,b,c,d)* $\in \Sigma_{\text{P}}$ – порівняння a з b, якщо a більше, то c – вкладена операція що виконується; якщо менша, то виконується d.

Інтерпритація конструктору. Сформуємо конструктивну систему з конструктору C_{P}, як елементної бази конструювання та алгорітмічного конструктору C_{A}, як моделі-виконавця конструювання.

$$\langle \text{C}_{\text{P}} = \langle M_{\text{P}}, \Sigma_{\text{P}}, \Lambda_{\text{P}} \rangle, \text{C}_{\text{A}} = \langle M_{\text{A}}, V_A, \Sigma_{\text{A}}, \Lambda_{\text{A}} \rangle \rangle \ {}_{\text{I}}\mapsto \langle \text{C}_{\text{PAI}} = \text{M}_{\text{I}}, \Sigma_{\text{I}}, \Lambda_{\text{I}} \rangle, \qquad (2.12)$$

де $V_A = \{A_i \mid_{X_i}^{Y_i}\}$ – множина утворюючих алгоритмів базової алгоритмічної структури, X_i та Y_i – множина можливих вхідних та вихідних даних алгоритму $A_i \mid_{X_i}^{Y_i}$, $M_A \supset \bigcup_{A_i \in V_A}(X(A_i) \cup Y(A_i))$ – носій алгоритмічної структури, Σ_{A} – множина операцій зв'язування алгоритмів, Λ_{A} –ІЗК, $\Omega(\text{C}_{\text{A}})$ – множина алгоритмів, що конструюються у C_{A}[4], $\text{M}_{\text{I}} = M_{\text{P}} \cup M_{\text{A}}$, $\Sigma_{\text{I}} = \Sigma_{\text{P}} \cup \Sigma_{\text{A}}$, $\Lambda_{\text{I}} = \Lambda_{\text{P}} \cup \Lambda_{\text{A}} \cup \left\{\left(A_0 \mid_{A_i, A_j}^{A_i \cdot A_j} \downarrow " \cdot "\right); \left(A_1 \mid_{l, s_i}^{l} \downarrow " \Rightarrow "\right); \left(A_2 \mid_{l, \Psi}^{l} \downarrow "\mid \Rightarrow "\right); (A_3 \mid_{\sigma}^{\Omega} \downarrow "|| \Rightarrow "\right)\left(A_4 \mid_{word, ends}^{pos\downarrow word} \downarrow " \odot "\right); \left(A_5 \mid_{word, ends}^{num\downarrow word} \downarrow " \circledast "\right); \left(A_6 \mid_{word, ends}^{gen\downarrow word} \downarrow " \circledcirc "\right); (A_7 \mid_{b}^{a} \downarrow " = "); \left(A_8 \mid_{a,b}^{c} \downarrow " + "\right); \left(A_9 \mid_{a,b}^{c/d} \downarrow " <> "\right)\right\}$

Структура C_{PAI} містить алгоритми виконання операцій:

- $A_0 \mid_{A_i, A_j}^{A_i \cdot A_j}$ – композиція алгоритмів, $A_i \cdot A_j$ – послідовне виконання алгоритму A_j після A_i;

- $A_1 \mid_{l, s_i}^{l}$ – підстановка, де l – поточна форма, s_i – правило, що виконуються;

- $A_2 \mid_{l, \Psi}^{l}$ – частковий вивід, де Ψ – множина правил продукції, що виконуються;

- $A_3|_{\sigma,\Psi}^{\Omega}$ – повний вивід, де σ – аксіома, Ψ – множина правил продукції, Ω – множина сформованих конструкцій;

- $A_4|_{word,ends}^{pos\downarrow word}$ – визначення для слова *word* його частини мови *pos;*

- $A_5|_{word,ends}^{num\downarrow word}$ – визначення для слова *word* його числу *num;*

- $A_6|_{word,ends}^{gen\downarrow word}$ – визначення для *word* його роду *gen;*

- $A_7|_{b}^{a}$ – присвоєння значення b змінній *a;*

- $A_8|_{a,b}^{c}$ – додавання $c = a + b$;

- $A_9|_{a,b}^{c/d}$ – виконання дії c або d за результатом порівняння a та b.

При **конкретизації** C_{PAI} параметризується:

$$C_{PAI} \mapsto_K C_{PAIK}(TT) = \langle M_K, \Sigma_K, \Lambda_K \rangle, \quad (2.13)$$

де $\Lambda_K \supset \Lambda_I \cup \{M_K = T_T \cup N\} \cup \Lambda_1$. TT – технічний текст, що подається для розбору.

У правилах підстановки ψ_i:<s_i, g_i> послідовність відношень підстановки s_i складається з відношення $s_{i,1}$ – розбір тексту TT, $s_{i,2}$ – формування набору слів $W_{i,j}$ з їх атрибутикою. Операції $g_{i,j}$ виконується після виконання $s_{i,1}$ та перед $s_{i,2}$.

Початкова умови конструювання: σ – нетермінал, з якого починається виведення та початкові значення max = 1, i=1 та j=1.

Умова завершення конструювання: тегований весь вхідний текст.

У першому правилі починається розбір тексту та формування першого елементу у тегованому тексті $W_{i,j}$

$$s_1 = \langle \sigma \to \eta, \qquad W_{i,j} \to \varepsilon \rangle$$

Розбір відбувається від одного символа до наступного з його перезаписом у $W_{i,j}$ для подальшого тегування.

$$s_2 = \langle \eta \to * \eta, \qquad W_{i,j} \to * W_{i,j} \rangle$$

При досягненні пробілу або синтаксичного знаку кінця речення виконується визначення встановлення тегів для слова та перехід до наступного.

$$s_3 = \langle \eta \to \underline{*} \eta, \qquad \varepsilon \rangle$$

Операції $\odot$, $\circledast$ та $\circledcirc$ в операціях над атрибутами встановлюють частину мови слова, його число та рід відповідно; відбувається перехід до наступного слова у реченні. Прапорець done для кожного зі слів встановлюється у 0 позиція, він буде в подальшому використаний для формування правил.

$$g_3 = \langle \odot\,(W_{i,j}, pos \downarrow W_{i,j}), \quad \circledast\,(W_{i,j}, num \downarrow W_{i,j}), \\ \circledcirc\,(W_{i,j}, gen \downarrow W_{i,j}), = (done \downarrow W_{i,j}, 0), +(j, 1, j)\ \rangle$$

Правило s_4 застосовується при досягнені кінця речення, та як і попереднє, виконується визначення встановлення тегів для слова та перехід до наступного речення. Виконується перехід до наступного слова. Вираховується та встановлюється довжина кожного речення та зберігається як атрибут його першого слова. Визначається довжина максимального речення, як атрибут найпершого слова у тексті. Разом з цим, для маркування закінчення речення на його кінцеву позицію буде записано $\perp(W_{i,j} = \perp)$. Це необхідно для коректної роботи наступного конструктора.

$$s_4 = \langle \eta \rightarrow^{\overline{*}} \sigma,\ \rangle$$

$$g_4 = \langle \odot\,(W_{i,j}, pos \downarrow W_{i,j}),\ \circledast\,(W_{i,j}, num \downarrow W_{i,j}),\ \circledcirc\,(W_{i,j}, gen \downarrow W_{i,j}),\ = \\ (l \downarrow W_{i,1}, j), <> (j, max \downarrow W_{1,1}, = (max \downarrow W_{1,1}, j), \varepsilon), = (done \downarrow W_{i,j}, 0), \\ +(j, 1, j), = (W_{i,j}, \perp), = (j, 1),\ +(i, 1, i)\ \rangle$$

Останнє правило використовується при досягненні кінця тексту та є завершальним. У атрибут am першого слова зберігається загальна кількість речень у тексті.

$$s_5 = \langle \eta \rightarrow \perp,\ \varepsilon \rangle$$

$$g_5 = \langle = (am \downarrow W_{1,1}, i) \rangle$$

Реалізація конструктора – формування мовних конструкцій з елементів її носія через виконання алгоритмів, що зв'язані з операціями сигнатури за правилами підстановки:

$$\mathrm{C_{PAIK}}\ _R\mapsto \bar{\Omega}(\mathrm{C_{PAIK}}(\mathrm{TT})) \tag{2.14}$$

де $\bar{\Omega}(\mathrm{C_{PAIK}}(\mathrm{TT})) = \Omega(\mathrm{C_{PAIK}}(\mathrm{TT}))$. $\bar{\Omega}$ – усі можливі результати роботи конструктора, однак, оскільки сформований конструктор побудований на основі конкретного тексту, отриманий оброблений текст Ω буде єдиним можливим варіантом. В результаті реалізації конструктора отримуємо оброблений текст з тегованими словами як $\Omega(\mathrm{C_{PAIK}}(TT))$.

Для прикладу візьмемо речення «Чорні ґрати розпанахали небо. Червоно-рожеве воно тянуло, манило.». Результат роботи конструктора буде мати вигляд:

$W_{1,1} = {}_{adj,plur,-}$Чорні; $W_{1,2} = {}_{n,plur,-}$ ґрати; $W_{1,3} = {}_{v,plur,-}$ розпанахал; $W_{1,4} = {}_{n,sing,n}$ небо; $W_{2,1} =$

${}_{adj,sing,n}$Червоно – рожеве; $W_{2,2} = {}_{pr,sing,n}$ воно; $W_{2,3} = {}_{v,sing,n}$ тянуло; $W_{2,4} = {}_{v,sing,n}$ манило

2.5.3 Конструктор-перетворювач тегованого тексту у множину формальних правил підстановок з вірогідністною мірою

Метою конструювання є побудова правила стохастичного конструктора, що формалізує синтактичну складову технічного тексту.

Початкова умови конструювання реалізація конструктору $\mathrm{C_P}$ – тегований текст Tg, отриманий в результаті реалізації конструктору $\mathrm{C_{PAIK}}$ – $\Omega(\mathrm{C_P}(\mathrm{TT}))$.

Умова завершення конструювання: кожне речення тегованого тексту перетворене на відповідну сукупність правил $\Omega(\mathrm{C_T}(R))$, що відбувається за умови $\tau 5 = true$, що встановлюється при досягненні останнього слова найдовшого речення у тексті. Це буде слугувати ознакою того, що всі інші слова у тексті вже були оброблені та процес будування правил завершено.

Конструктор має наступну спеціалізацію:

$$C = \langle M, \Sigma, \Lambda\rangle \quad _S \mapsto C_T(Tg) = \langle M_T, \Sigma_T, \Lambda_T\rangle \tag{2.15}$$

де M_T – носій, що включає тегований текст Tg, Σ_T – операції та відносини на елементах M_T та аксіоматика Λ_T.

Операція $*$ (r, a, b) – перевірка на відповідність атрибутів pos↲ a, num↲ a, gen↲ a елемента a атрибутів pos↲ b, num↲ b, gen↲ b елемента b, де a та b це теговані слова. При повній відповідності результатом є 1, інакше – 0.

Операція $\&(\mathrm{y}, \mathrm{x1}, \mathrm{x2})$ – логічне та з необмеженою кількістю оперантів y = x1 та x2 та …;

Операція циклу $\circ$ (a, c) – a – умова, c – операція, що виконується поки умова є дійсною;

Операція $-(\mathrm{c}, \mathrm{a}, \mathrm{b})$ – тотожно c = a - b у інфіксній формі;

Операція : $(\mathrm{c}, \mathrm{a}, \mathrm{b})$ – тотожно c = a : b у інфіксній формі, ділення дійсних чисел;

Операція $\leq$(r,a,b) – порівняння a $\leq$ b з подальшим збереженням результату у r.

Інтерпретуємо конструктор C_T за допомогою того що й раніше алгоритмічного конструктора C_A:

$$\langle C_T, \mathrm{C_A} \rangle \ {}_{\mathrm{I}}\mapsto \langle \mathrm{C}_T = \mathrm{M_{TI}}, \Sigma_{\mathrm{TI}}, \Lambda_{\mathrm{TI}} \rangle, \tag{2.16}$$

$\mathrm{M_{TI}}$ – алгоритмічна структури для формування стохастичного конструктора з тегованого тексту, Σ_{TI} – операції зв'язування алгоритмів, $\Lambda_{\mathrm{TI}} \supset \Lambda_{\mathrm{I}} \cup \Lambda_1 \cup \Lambda_2$.

$\Lambda_2 = \{ (A_{10}|_{a,b}^{r} \text{↲ " * "}); (A_{11}|_{a,b}^{r} \text{↲ " △ "}); (A_{12}|_{a,b}^{c} \text{↲ " − "}); (A_{13}|_{a,b}^{c} \text{↲ ": "}); (A_{14}|_{a,b}^{r} \text{↲ " ≤ "}). \}$.

Структура C_{TAI} включає наступні алгоритми:

- $A_0, A_1, A_2, A_3, A_7, A_8, A_9$ – аналогічні алгоритми конструктору $\mathrm{C_{PAI}}$;
- $A_{10}|_{a,b}^{r}$ – порівняння a та b на ідентичність;
- $A_{11}|_{a,b}^{r}$ – логічне «так»;
- $A_{12}|_{a,b}^{\mathrm{c}}$ – віднімання чисел;
- $A_{13}|_{a,b}^{\mathrm{c}}$ – ділення дійсних чисел;
- $A_{14}|_{a,b}^{r}$ – порівняння a та b.

Конкретизації $\mathrm{C_T}$:

$$C_T \mapsto_{\mathrm{K}} C_T(\mathrm{C}_P(Tg)) = \langle \mathrm{M_K}, \Sigma_{\mathrm{K}}, \Lambda_{\mathrm{K}} \rangle, \tag{2.17}$$

де $\Lambda_K \supset \Lambda_I$, $\Lambda_K \supset \{M_K = T \cup N\}$, до терміналів T належать усі слова $W_{m,j}$ з позначенням їх місця j у реченні m, $\alpha_{k,j}$, що являє собою нетермінал правила, що будується, σ –початковий нетермінал та побудоване правило ω_k, що у своїх атрибутах буде мати ліву частину правила L, праву частину R та вірогідність його спрацювання для даного тексту prob. До нетерміналів N: τ_i – атрибут доступности правила.

Для кожної частини мови прораховується ймовірність (prob) її появи у певному місці певного речення у цьому тексті. Імовірність появи певної частини мови в досліджуваній послідовності дозволить більш точно вловити індивідуальний стиль письма, характерний кожному з авторів, що досліджуються.

Імовірність виведення всього речення визначається як добуток ймовірностей, які у ньому послідовностей частин мови. Отриманий конструктор породжуватиме мову, характерну для оброблюваного тексту та структурно подібних текстів певного автора.

Таким чином, кожне з речень представленого тексту буде представлено у вигляді ланцюжка правил, що відображатиме послідовність використаних частин мови та вірогідності їх появи саме у представленій послідовності.

Початкові умови: початкова форма $W_{1,1}$ – перше слово у тексті, де i = 1, n = 2 номера речень, j = 1, m = 2 номера слів у них. t = 0 – кількість співпадінь з обраною парою за параметрами у послідовність з двох слів, k = 1 – номер правила, що будується. $\tau 1 = true, \tau 2 = false \ldots \tau 5 = false$ – умови виконання правил: якщо $true$, то доступно до використання, якщо $false$ – ні. *idone* = 1, *jdone* = 1 змінні що дорівнюють номеру унікального елементу у шарі та попередньому шарі відповідно, *u = false* – прапорець для відмітки вже побудованих правил.

Розбір починається з опрацювання першого шару (перших слів у реченнях тексту) та пошуку співпадіння за атрибутами серед них

$$s_1 = \langle W_{i,j\ \tau 1} \rightarrow W_{n,j}, \varepsilon \rangle,$$

$$g_{1,1} = \langle == (0, \text{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}), * (\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}) \rangle.$$

Шукається співпадіння слів з однаковою атрибутикою у поточному шарі. Якщо співпадіння з поточним словом цього поточного шару знайдено, та для цього слова не було виявлено співпадіння раніше, збільшуємо загальну кількість подібних до шуканого (t) та переходимо до наступного речення збільшуючи n

$$g_{1,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2}), <> (y, 0, +(t, 1, t), \varepsilon\ ), +(n, 1, n) \rangle.$$

Друге правило використовується при досягненні кінця речення, у цьому випадку розрахунків не відбувається, лише збільшується номер речення n у для переходу до наступного слова у шарі

$$s_2 = \langle W_{i,j}\ {}_{\tau 1}\!\rightarrow W_{n,j}\ \perp, \varepsilon \rangle,$$

$$g_2 = \langle +(n, 1, n) \rangle.$$

Третє правило застосовується при неможливості до наступного слова у шарі через досягнення його кінця

$$s_3 = \langle W_{i,j}\ {}_{\tau 1}\!\rightarrow \perp, \varepsilon \rangle.$$

За цієї умови використовується підрахунок вірогідності появи обраної послідовності у тексті, правила $s_1 - s_3$ стають недосяжними, а правила $s_4 - s_7$ стають доступними для опрацювання

$$g_3 = \langle -(n, n, 1), : (prob, t, n), = (\tau 1, false), = (\tau 2, true), = (n, 1) \rangle.$$

Наступні правила відповідають за формування правил для перших слів кожного з речень при повторному проході усіх відповідних слів

$$s_4 = \langle W_{i,j}\ {}_{\tau 2}\!\rightarrow W_{n,j},\ \omega_k \rangle,$$

$$g_{4,1} = \langle == \left(0, \mathrm{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}\right), * \left(\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}\right) \rangle.$$

Якщо співпадає атрибутика слів будується нове правило ω_k. У ліву та праву його частину записуються відповідно σ та $W_{i,j}\alpha_{i,j}$, де $\alpha_{i,j}$ – нетермінал

новосформованих правил, та записується вірогідність його спрацювання для тексту $prob$

$$g_{4,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2},),$$
$$<> (y, 0, \cdot \left(= \left(L \downarrow \omega_{i,j}, \sigma\right), = \left(R \downarrow \omega_{i,j}, W_{i,j}\alpha_{i,j}\right), = \left(prob \downarrow \omega_{i,j}, prob\right)\right), \varepsilon)\rangle.$$

Після формування правила встановлюється прапорець наявності хоча б одного правила на цьому шарі $u = true$, правило отримує індекс унікальності у шарі $idone$ та конструктор переходить до наступного речення

$$g_{4,4} = \langle +(n, 1, n), = \left(done \downarrow \mathrm{W_{i,j}}, idone\right), = (u, true)\rangle.$$

Наступне правило аналогічне правилу s_2, не робить розрахунків та відповідає за збільшення номеру речення n у для подальшого руху по шару

$$s_5 = \langle W_{i,j} \ _{\tau 2}\rightarrow W_{n,j} \perp, \varepsilon\rangle,$$

$$g_5 = \langle +(n, 1, n)\rangle.$$

Якщо у реченні лише одне слово, використовується правило s_6

$$s_6 = \langle W_{i,j} \perp_{\tau 2}\rightarrow W_{n,j}, \omega_k \rangle.$$

Для формування правила у цьому випадку будуть проводитись наступні перевірки: чи включене слово до іншого правила done, чи співпадають відповідні атрибути слів

$$g_{6,1} = \langle == \left(0, \mathrm{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}\right), * \left(\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}\right)\rangle.$$

Якщо речення не перше формуємо відповідне правило ω_k. У ліву його частину записуються σ відповідно та лише $W_{i,m}$ у праву

$$g_{6,3} = \langle \&(\mathrm{z}, \mathrm{x1}, \mathrm{x2}, \mathrm{x3}), <> \left(z, 0, \cdot (= \left(L \downarrow \omega_{i,j}, \sigma\right), = \left(R \downarrow \omega_{i,j}, W_{i,j}\right), = \left(prob \downarrow \omega_{i,j}, prob\right)\right), \varepsilon \right)\rangle.$$

І далі так само як у $g_{4,4}$ – встановлюється прапорець його створення $u = true$, правило отримує індекс унікальності у шарі $idone$ та виконавець переходить до наступного речення

$$g_{6,4} = \langle +(n, 1, n), = (done \downarrow \mathrm{W_{i,j}}, idone), = (u, true) \rangle.$$

І при досягненні кінця спрацьовує правило s_7

$$s_7 = \langle W_{i,j}\ _{\tau 2} \rightarrow \perp, \varepsilon \rangle.$$

Досягнення кінця шару означає закінчення формування правила та перехід до формування іншого. Для цього змінюються прапореці $\tau 2 = false$ та $\tau 1 = true$, що дозволить закрити правила $s_4 - s_7$ та відкрити правила $s_1 - s_3$ для пошуку інших співпадінь та їх підрахунку. Для відображення роботи з іншим правилом у шарі номер унікальності правила для шару $idone$ збільшується

$$g_{7,1} = \langle == (u, true, y), <> (y, 0, \cdot \big(= (\tau 2, false), = (\tau 1, true), +(idone, 1, idone), = (t, 0)\big), \varepsilon \rangle.$$

Якщо роботу з шаром завершено і для усіх слів у ньому були сформовані правила, конструктор переходить на наступний шар починаючи знов з першого речення $i = 1$ шукати співпадіння. Розрахунок унікальності правил у шарі також починається з початку $idone = 1$. За умови досягнення кінцевого шару (оброблено останнє слово у найдовшому реченні $\mathrm{W}_{i,max}$) роботу виконавця з першим шаром буде завершено $\tau 3 = true, \tau 1 = false, \tau 2 = false$

$$g_{7,2} = \langle == (u, false, y), <> (y, 0, \cdot (= (i, 1), = (idone, 1), = (t, 0), = (\tau 3, true), = (\tau 2, false), = (\tau 1, false)), \varepsilon \rangle.$$

Для продовження формування правил з тегованого тексту будемо працювати з послідовними парами слів у кожному реченні. Перехід від слова до слова відбувається не вздовж речення, а відповідно номеру слів у них. Таким чином конструктором розглядаються пара слів поспіль у реченні

$$s_8 = \langle W_{i,j} W_{i,m}\ _{\tau 3} \rightarrow W_{n,j} W_{n,m}, \varepsilon \rangle.$$

Для врахування існуючої пари слів до подібних потрібно перевірити наступні параметри: слова ще не були опрацьовані; атрибути обраної послідовності двох слів (частина мови, рід та число) співпадають з відповідними за номером словами у наступному реченні, попередній ланцюжок слів також повинен співпадати, що перевіряється за допомогою jdone

$$g_{8,1} = \langle == \left(0, \text{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}\right), * \left(\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}\right), * \left(\mathrm{W_{i,m}}, \mathrm{W_{n,m}}, \mathrm{x3}\right), -(\mathrm{k}, \mathrm{j}, 1), <$$
$$> \left(\mathrm{j}, 1, == \left(\text{ done} \downarrow \mathrm{W_{i,k,}}\text{jdone}, \mathrm{x4}\right), \mathrm{x4} = \text{true}\right)\rangle.$$

При співпадінні атрибутів пара зараховується в загальну кількість подібних послідовностей та збільшується значення t та значення n для переходу до наступного речення у шарі

$$g_{8,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2}, \mathrm{x3}, \mathrm{x4}), <> (y, 0, +(t, 1, t), \varepsilon), +(n, 1, n)\rangle.$$

Наступне правило спрацьовує при досягненні кінці речення, збільшується номер речення n у для подальшого перегляду слів у шарі

$$s_9 = \langle W_{i,j} W_{i,m} \ {}_{\tau 3}\!\rightarrow W_{n,j} \ \perp, \varepsilon\rangle,$$

$$g_9 = \langle +(n, 1, n)\rangle.$$

Подальше правило виконується при неможливості руху далі по реченнях через досягнення кінця шару. За цієї умови відбувається підрахунок вірогідності появи обраної послідовності у тексті, правила $s_8 - s_{10}$ стають недосяжними, а правила $s_{11} - s_{14}$ стають доступними для опрацювання

$$s_{10} = \langle W_{i,j} \ {}_{\tau 3}\!\rightarrow \perp, \varepsilon\rangle,$$

$$g_{10} = \langle -(n, n, 1), : (prob, t, n), = (\tau 3, false), = (\tau 4, true), = (i, 1), = (n, 1)\rangle.$$

Наступним кроком потрібно ще раз переглянути поточний шар та сформувати правила ω_k

$$s_{11} = \langle W_{i,j} W_{i,m} \ {}_{\tau 4}\!\rightarrow W_{n,j} W_{n,m}, \ \omega_k\rangle.$$

Для формування відповідного правила повторюється перевірка з першого правила та додатково перевіряємо чи є слово першим у реченні (x4) для корректного формування початкових правил

$$g_{11,1} = \langle == (0, \text{done} \downarrow \text{W}_{\text{i,j}}, \text{x1}), * (\text{W}_{\text{i,j}}, \text{W}_{\text{n,j}}, \text{x2}), * (\text{W}_{\text{i,m}}, \text{W}_{\text{n,m}}, \text{x3}), -(\text{k}, \text{j}, 1), <>$$
$$(\text{j}, 1, == (\text{done} \downarrow \text{W}_{\text{i,k,}}\text{jdone}, \text{x4}), \text{x4} = \text{true})\rangle.$$

Якщо все співпадає та слово не є першим в реченні будується нове правило ω_k. У ліву частину правила записуються $\alpha_{i,j}$, праву його частину $W_{i,m}\alpha_{i,m}$, де $\alpha_{i,j}$ – нетермінал новосформованих правил, та записується вірогідність його спрацювання для тексту $prob$.

$$g_{11,2} = \langle \&(\text{y}, \text{x1}, \text{x2}, \text{x3}, \text{x4}),$$
$$<> (y, 0, \cdot (= (L \downarrow \omega_{i,j}, \alpha_{i,j}), = (R \downarrow \omega_{i,j}, W_{i,m}\alpha_{i,m}), = (prob \downarrow \omega_{i,j}, prob)), \varepsilon)\rangle.$$

Після формування правила встановлюється прапорець його створення $u = true$, правило отримує індекс унікальності у шарі $idone$ та конструктор переходить до наступного речення

$$g_{11,3} = \langle +(n, 1, n), = (done \downarrow \text{W}_{\text{i,j}}, idone), = (u, true)\rangle.$$

Наступне правило аналогічне правилу s_9, не робить розрахунків та відповідає за збільшення номеру речення n у для подальшого руху по шару

$$s_{12} = \langle W_{i,j} W_{i,m} \ _{\tau 4}\rightarrow W_{n,j} \perp, \varepsilon\rangle,$$

$$g_{12} = \langle +(n, 1, n)\rangle.$$

Якщо у шарі знаходиться останнє слово у реченні, спрацьовує правило s_6

$$s_{13} = \langle W_{i,j} \perp_{\tau 4}\rightarrow W_{n,j}, \omega_k \rangle.$$

Для формування правила у цьому випадку будуть проводитись наступні перевірки: чи включене слово до іншого правила done, чи співпадають відповідні атрибути слів, чи є слово першим у реченні і, якщо ні, чи співпадає попередній ланцюжок.

$$g_{13,1} = \langle == (0, \text{done} \downarrow \text{W}_{\text{i,j}}, \text{x1}), * (\text{W}_{\text{i,j}}, \text{W}_{\text{n,j}}, \text{x2}), <> (\text{j}, 1, == (\text{done} \downarrow \text{W}_{\text{i,j}-1}, \text{jdone}, \text{x3}), \text{x3} = \text{true})\rangle.$$

Якщо речення не перше формуємо відповідне правило ω_k. У ліву його частину записуються відповідо $\alpha_{i,j}$ та лише $W_{i,m}$ у праву, де $\alpha_{i,j}$ – нетермінал новосформованих правил, та записується вірогідність його спрацювання для тексту $prob$

$$g_{13,2} = \langle \&(\text{y}, \text{x1}, \text{x2}, \text{x3}), <> (y, 0, \cdot (= (L \downarrow \omega_{i,j}, \alpha_{i,j}), = (R \downarrow \omega_{i,j}, W_{i,j}), = (prob \downarrow \omega_{i,j}, prob)), \varepsilon)\rangle.$$

Далі так само як у $g_{11,3}$ – встановлюється прапорець його створення $u = true$, правило отримує індекс унікальності у шарі $idone$ та конструктор переходить до наступного речення

$$g_{13,3} = \langle +(n, 1, n), = (done \downarrow \text{W}_{\text{i,j}}, idone), = (u, true)\rangle.$$

При досягненні кінця спрацьовує правило s_7

$$s_{14} = \langle W_{i,j\ \tau 4} \rightarrow \perp, \varepsilon\rangle.$$

Досягнення кінця шару означає закінчення формування правила та перехід до формування іншого. Для цього змінюються прапореці $\tau 4 = false$ та $\tau 3 = true$, що дозволить закрити правила $s_{11} - s_{14}$ та відкрити правила $s_8 - s_{10}$ для пошуку інших співпадінь та їх підрахунку. Для відображення роботи з іншим правилом у шарі номер унікальності правила для шару $idone$ збільшується

$$g_{14,1} = \langle = (\tau 4, false), = (\tau 3, true), +(idone, 1, idone), = (t, 0), \leq (r, j, l \downarrow \text{W}_{\text{i,j}})\rangle.$$

Процедура підрахунку співпадінь та розрахунок вірогідності їх появи для будування на її основі правил відбувається доки усі слова у шарі не будуть оброблені. Для роботи з усіма ланцюжками при кожному проході шару збільшується індекс унікальності $jdone$ для перевірки умови розрахунків $g_{1,1}$

$$g_{14,2} = \langle <> (am \downarrow \mathrm{W}_{1,1}, i, \circ (r, \cdot (= (i, 1, i) , == (u, true, 1) < > (1,0, +(jdone, 1, jdone))) \rangle.$$

Якщо роботу з шаром завершено і для усіх слів у ньому були сформовані правила, конструктор переходить до наступного шару збільшуючи j та починаючи знов з першого речення $i = 1$ шукати співпадіння. Розрахунок унікальності правил у шарі також починається з початку $idone = 1$ та $jdone = 1$. За умови досягнення кінцевого шару (оброблено останнє слово у найдовшому реченні $\mathrm{W}_{i,\mathrm{m}ax}$) роботу конструктора буде завершено $\tau 5 = true$

$$g_{14,3} = \langle <> (max \downarrow \mathrm{W}_{1,1}, j, \cdot (+(j, 1, j), = (i, 1), = (idone, 1), = (jdone, 1)), \cdot (= (\tau 5, true), = (\tau 3, true))) \rangle.$$

У результаті роботи конструктора-перетворювача з $\Omega(\mathrm{C}_{\mathrm{PAIK}}\ (TT))$ отримуємо множину правил, що відображає стиль мови автора у відповідному тексті $\Omega(\mathrm{C}_{\mathrm{T}}(R))$.

Реалізація структури – формування мовних конструкцій з елементів її носія через виконання алгоритмів, що зв'язані з операціями сігнатури за правилами аксіономатики:

$$\mathrm{C}_{\mathrm{P\,K}}\ _R \mapsto \bar{\Omega}(C_{P\,K}), \tag{2.18}$$

де $\bar{\Omega}(C_{P\,K}) \subset \Omega(\mathrm{C}_{\mathrm{P\,K}})$.

Для прикладу візьмемо речення, що мають вигляд:

«Ми були дуже схожі.
Я любила читати книжки.
А ти захоплювався виставами.
Але..
Між нами було й багато різниці».

Тегований текст для цього прикладу:

$W_{1,1} = {}_{pron,plur}$Ми $W_{1,2} =_{v,plur}$ були $W_{1,3} =_{adv,sing}$ дуже $W_{1,4} =_{adj,plur}$ схожі

$W_{2,1} = {}_{pron,sing}$Я $W_{2,2} =_{v,sing}$ любила $W_{2,3} =_{v,sing}$ читати $W_{2,4} =_{n,plur}$ книжки

$$W_{3,1} =$$

$_{conj}$А $W_{3,2} =_{pron,sing}$ ти $W_{3,3} =_{v,sing}$ захоплювався $W_{3,4} =_{n,plur}$ виставами

$$W_{4,1} = {}_{pron}\text{Але}$$

$$W_{5,1} = {}_{conj}\text{Між}\ W_{5,2} =_{pron,plur} \text{нами}\ W_{5,3} =_{v,sing}$$

було $W_{5,4} =_{conj}$ й $W_{5,5} =_{adj,sing}$ багато $W_{5,6} =_{adj,plur}$ різного.

Результат роботи конструктора буде представлений у вигляді відповідних правил:

$$\sigma \overset{0.2}{\rightarrow} W_{1,1}\alpha_{1,1};\ \alpha_{1,1} \overset{0.2}{\rightarrow} W_{1,2}\alpha_{1,2};\ \alpha_{1,2} \overset{0.2}{\rightarrow} W_{1,3}\alpha_{1,3};\ \alpha_{1,3} \overset{0.2}{\rightarrow} W_{1,4};$$

$$\sigma \overset{0.2}{\rightarrow} W_{2,1}\alpha_{1,1};\ \alpha_{2,1} \overset{0.2}{\rightarrow} W_{2,2}\alpha_{2,2};\ \alpha_{2,2} \overset{0.6}{\rightarrow} W_{2,3}\alpha_{2,3};\ \alpha_{2,3} \overset{0.4}{\rightarrow} W_{2,4};$$

$$\sigma \overset{0.4}{\rightarrow} W_{3,1}\alpha_{3,1};\ \alpha_{3,1} \overset{0.2}{\rightarrow} W_{3,2}\alpha_{3,2};\ \alpha_{3,2} \overset{0.6}{\rightarrow} W_{3,3}\alpha_{3,3};\ \alpha_{3,3} \overset{0.4}{\rightarrow} W_{3,4};$$

$$\sigma \overset{0.2}{\rightarrow} W_{4,1};$$

$$\sigma \overset{0.4}{\rightarrow} W_{5,1}\alpha_{5,1};\ \alpha_{5,1} \overset{0.2}{\rightarrow} W_{5,2}\alpha_{5,2};$$

$$\alpha_{5,2} \overset{0.6}{\rightarrow} W_{5,3}\alpha_{5,3};\ \alpha_{5,3} \overset{0.2}{\rightarrow} W_{5,4}\alpha_{5,4};\ \alpha_{5,4} \overset{1}{\rightarrow} W_{5,5}\alpha_{5,5};\ \alpha_{5,5} \overset{1}{\rightarrow} W_{5,6}\ .$$

2.5.4 Конструктор-вимірювач ступеню подібності двох текстів

Для встановлення ступеня схожості текстів за синтаксичним стилем мови автора проводиться порівняння моделей тексту за допомогою конструктора-вимірювача.

Метою конструювання є встановлення ступеня схожості текстів порівнянням стохастичних конструкторів, побудованих за їх синтаксичною структурою.

Початкова умови конструювання моделі двох текстів у вигляді множини правил підстановки з вірогідністю його спрацювання $\Omega(\mathrm{C_T}(R_1))$ та $\Omega(\mathrm{C_T}(R_2))$, які представляють текст певних технічних робіт $\Omega(\mathrm{C_{PAIK}}(TT_1))$ та $\Omega(\mathrm{C_{PAIK}}(TT_2))$, що є результатом виконання попередніх конструкторів.

Умова завершення конструювання: $\tau 3 = true$, отривання числа від 0 до 1, що відображає подібність двох робіт після порівняння усіх правил в двох моделях тексту.

Конструктор має наступну спеціалізацію:

$$C = \langle M, \Sigma, \Lambda \rangle \ _S \mapsto C_{\mathrm{E}} = \langle M_{\mathrm{E}}, \Sigma_{\mathrm{E}}, \Lambda_{\mathrm{E}} \rangle, \quad (2.19)$$

де M_{E} – носій, який включає множину правил, що описують мову автора у певному тексті R_i, Σ_E – операції та відносини на елементах M_E та ІЗК Λ_E.

Інтерпретуємо структуру C_E за допомогою алгоритмічної структури C_{A}:

$$\langle C_E, \mathrm{C_A} \rangle \ _{\mathrm{I}} \mapsto \langle \mathrm{C}_E = \mathrm{M_{PI}}, \Sigma_{\mathrm{PI}}, \Lambda_{\mathrm{PI}} \rangle, \quad (2.20)$$

де $V_A = \{A_i^0|_{X_i}^{Y_i}\}$ – множина утворюючих алгоритмів базової алгоритмічної структури, $\mathrm{X_i}$ та $\mathrm{Y_i}$ – множина визначень та значень алгоритму $A_i^0|_{X_i}^{Y_i}$, $M_A = \bigcup_{A_i^0 \in V_A}(X(A_i^0) \cup Y(A_i^0))$ – носій алгоритмічної структури, Σ_{I} – множина операцій зв'язування алгоритмів, Λ_{I} –аксіоматика алгоритмічної структури, $\Omega(\mathrm{C_A})$ – множина алгоритмів, що конструюються у $\mathrm{C_A}$.

Надалі представлені операція над атрибутами.

Операція min(m, a, b) порівнює числа a та b, та зберігає найменше у m;

Операція –(c, a, b) – віднімання *c = a – b;*

Операція *(c, a, b) – множення *c = a * b;*

$\mathrm{M_{PI}}$ – алгоритмічна структури для порівняння правил, Σ_{PI} – операції зв'язування алгоритмів, $\Lambda_{\mathrm{TI}} \supset \Lambda_{\mathrm{I}} \cup \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$.

$\Lambda_3 = \{ (A_{15}|_{a,b}^m \downarrow "min"); (A_{16}|_{a,b}^m \downarrow " - "); (A_{17}|_{a,b}^m \downarrow " * "); (A_{18}|_{a,b}^m \downarrow "max") \}$.

Структура $\mathrm{C_{PAI}}$ включає наступні алгоритми:

- $A_0, A_1, A_2, A_3, A_6, A_7, A_8, A_9, A_{10}$ – аналогічні алгоритми структури C_{PAI} та $\mathrm{C_{TAI}}$;

- $A_{15}|_{a,b}^m$ – пошук мінімума серед чисел a та b;

- $A_{16}|_{a,b}^{\mathrm{c}}$ – віднімання *c = a – b;*

- $A_{17}|_{a,b}^{\mathrm{c}}$ – множення *c = a * b;*

- $A_{18}|_{a,b}^m$ – пошук максимума серед чисел a та b;

Конкретизації $\mathrm{C_T}$:

$$C_E \ _K \longmapsto \ C_E(\Omega(\mathrm{CT}\ (R_1)), \Omega(\mathrm{CT}\ (R_2))) = \langle \mathrm{M_K}, \Sigma_{\mathrm{K}}, \Lambda_{\mathrm{K}} \rangle, \quad (2.21)$$

де $\Lambda_{\mathrm{K}} \supset \Lambda_{\mathrm{I}}$, $\Lambda_K \supset \{M_K = T_T \cup N\}$. до терміналів T належать усі слова у правилах обох конструкторів, що порівнюються ω та $\dot{\omega}$, до нетерміналів N – допоміжний символ ι.

У термінах конструктивно-продукційного моделювання процес порівняння сукупності правил для формування двох текстів (T_1 та T_2 відповідно) та отримання кінцевого значення їх подібності.

Перше правило розпочинає порівняння правил двох конструкторів $\Omega(C_T(R_1))$ та $\Omega(C_T(R_2))$, що описують два тексти, що досліджуються на їх подібність, i =1, j=1.

При існування однакових правил або правил ступінь їх статистичної структурної подоби визначатиметься як добуток мінімальної різниці ймовірностей застосування відповідного правила

$$\rho(\vartheta_i, \dot{\vartheta}_j) = \prod_{m=1}^{l} min(prob_m - \dot{prob}_m), \quad (2.22)$$

де ϑ_i – і-те речення з тексту T_1 та $\dot{\vartheta}_j$ – j-те речення з тексту T_2.

Ступінь статистичної структурної подоби текстів T_1 та T_2:

$$\rho(T_1, T_2) = \sum_{i=1}^{N} \rho(\vartheta_i, \dot{\vartheta}_j), \quad (2.23)$$

Початкові умови: $rule$ = 1, i = m = j = n = 1, де i та m – номери ланцюжків (речень) у тексті, j та n – номери правил у ланцюжках. $max \downarrow \omega_{i,1}$, де max = 0, де мах – добуток різниці вірогідностей. $max\_ch \downarrow \omega_{i,1}$, max_ch = 0, максимальна довжина ланцюжка, res = 0 – загальна подібність двох текстів, $k = n + 1, h = j + 1$, це наступні правила у ланцюжку відносно j та n відповідно. Та прапореці для спрацьовування s_1 та s_2 $\tau 1 = true$, $\tau 2 = false$, а також прапорець для завершення порівняння $\tau 3 = false$.

Перше правило використовується для порівняння перших правил в усіх ланцюжках тексту

$$s_1 = \langle \sigma_{\tau 1} \rightarrow \vartheta_{i,1}; \sigma_{\tau 1} \rightarrow \vartheta_{m,1} \rangle.$$

Для кожного правила при співпадінні їх правих частин та за умови що довжина

ланцюжка складає лише одне правило (тобто речення складається лише з одного слова)

$g_{1,1} = \langle * \left(R \downarrow \vartheta_{i,1}, R \downarrow \dot{\vartheta}_{m,1}, x1\right), == \left(l \downarrow W_{i,1} \downarrow \vartheta_{i,1}, 1, x2\right), == \left(l \downarrow W_{m,1} \downarrow \vartheta_{m,1}, 1, x3\right), \&(y, x1, x2, x3)\rangle.$

За умови виконання усіх умов розраховується добуток різниці їх вірогідностей, та результат зберігається у першому елементі ланцюжка. І поки не досягається кінець другого тексту, добутки підсумовуються у res. Якщо закінчуються ланцюжки з першого тексту – перше правило закривається та відкривається друге

$$g_{1,3} = \langle <> (y, 0, \cdot (* \left(max \downarrow \vartheta_{i,1}, min\left(-\left(r, prob \downarrow \vartheta_{i,h}, prob \downarrow \dot{\vartheta}_{m,k}\right)\right)\right),$$
$$<> (max \downarrow W_{1,1}$$
$$\downarrow \vartheta_{m,1}, m, \varepsilon, \cdot \left(+\left(res, max \downarrow \vartheta_{i,1}, r\right), +(i, 1, i), = (m, 1)\right),$$
$$<> (max \downarrow W_{1,1} \downarrow \vartheta_{i,1}, i, \varepsilon, \cdot \left(= (\tau 1, false), = (\tau 2, true)\right))), \varepsilon) \rangle.$$

У другому правилі послідовно перебираються усі ланцюжки, довші за одне правило, з просуванням по їх довжині для обох текстів, що досліджуються. Виконується перегляд всіх правил другого тексту (m змінюється від 1 до кінця тексту). Для кожного речення виконується послідовний перегляд всіх правил

$$s_2 = \langle \vartheta_{i,j\ \tau 2} \rightarrow\ \vartheta_{i,h},\ \dot{\vartheta}_{m,n\ \tau 2} \rightarrow \dot{\vartheta}_{m,k} \rangle.$$

Для початку роботи та розрахунку подібностей порівнюються праві частини перших правил в обох текстах і виконуємо операції над атрибутами

$$g_{2,1} = \langle * \left(R \downarrow \vartheta_{i,j}, R \downarrow \dot{\vartheta}_{m,n}, x1\right), <> \left(l \downarrow W_{i,1} \downarrow \vartheta_{i,j}, j, = (x2, true)\right), = (x2, false)\right), <> (l \downarrow W_{m,1} \downarrow \dot{\vartheta}_{m,n}, n, = (x3, true), = (x3, false)), == (j, 1, x4)\rangle.$$

В разі виконання усіх умов, обробляється перше правило у ланцюжку: ведеться підрахунок довжини ланцюжка, що співпадає ch, знаходиться добуток різниці вірогідностей правил з обох текстів sim та у першому елементу ланцюжка

зберігається максимальна довжина співпадаючого ланцюжка та результат обчислення їх співпадіння

$$g_{2,2} = \langle \&(y, x1, x2, x3, x4), <> (y, 0, \cdot \left(+(ch \lrcorner \vartheta_{i,1}, 1, ch \lrcorner \vartheta_{i,1}\right), *$$
$$\left(rule, min\left(-\left(r, prob \lrcorner \vartheta_{i,j}, prob \lrcorner \dot{\vartheta}_{m,n}\right)\right)\right), = \left(sim \lrcorner \vartheta_{i,1}, rule\right), <>$$
$$\left(ch \lrcorner \vartheta_{i,1}, maxch \lrcorner \vartheta_{i,1}, \cdot \left(= \left(maxch \lrcorner \vartheta_{i,1}, ch \lrcorner \vartheta_{i,1}\right), = \left(max \lrcorner \vartheta_{i,1}, sim \lrcorner \right.\right.\right.$$
$$\left.\left.\left. \vartheta_{i,1}\right)\right), +(j, 1, j), +(n, 1, n)), \cdot (= (j, 1), = (n, 1), +(m, 1, m)), \varepsilon) \rangle.$$

Потім за тих же умов оброблюються усі наступні ланцюжки та правила у них

$$g_{2,3} = \langle * \left(R \lrcorner \vartheta_{i,h}, R \lrcorner \dot{\vartheta}_{m,k}, x1\right), <$$
$$> \left(l \lrcorner W_{i,1} \lrcorner \vartheta_{i,h}, h, = (x2, true)), = (x2, false)\right),$$
$$<> \left(l \lrcorner W_{m,1} \lrcorner \dot{\vartheta}_{m,n}, k, = (x3, true), = (x3, false)\right), \&(y, x1, x2, x3)\rangle.$$

Якщо ланцюжок співпадінь обірвався – починається порівняння правил спочатку 2-го тексту вже для наступного ланцюжка правил першого тексту. Якщо ланцюжок закінчився – виконується перехід до наступного та так само перевіряється кожне з правил в обох текстах на співпадіння. У разі закінчення правил у тексті закриваємо можливість виконання другого правила $\tau 2 = false$ та закінчуємо розрахунки за допомогою прапорця $\tau 3 = true$

$$g_{2,4} = \langle <> (y, 0, \cdot \left(+(ch \lrcorner \vartheta_{i,1}, 1, ch \lrcorner \vartheta_{i,1}), * \left(rule, min\left(-\left(r, prob \lrcorner \right.\right.\right.\right.$$
$$\left.\left.\left.\left. \vartheta_{i,h}, prob \lrcorner \dot{\vartheta}_{m,k}\right)\right)\right)\right), = \left(sim \lrcorner \vartheta_{i,1}, rule\right), +(h, 1, h), +(k, 1, k)), <> (max \lrcorner W_{1,1} \lrcorner$$
$$\vartheta_{m,1}, m, \varepsilon, \cdot \left(+\left(res, max \lrcorner \vartheta_{i,1}, r\right), +(i, 1, i), = (n, 1), = (m, 1), +(k, n, 1), = \right.$$
$$\left.(j, 1), +(h, j, 1)\right), <> (max \lrcorner W_{1,1} \lrcorner \vartheta_{i,1}, i, \varepsilon, \cdot \left(= (\tau 2, false), = (\tau 3, true)\right)\rangle.$$

Відмітимо, що $\rho(T_1, T_2) = \rho(T_2, T_1)$ $\rho(T_1, T_1) = 1$ повне співпадіння, $\rho(T_1, T_2) = 0$ – якщо у текстах T_1 та T_2 немає речень однакової структури.

Реалізація структури – формування мовних конструкцій з елементів її носія через виконання алгоритмів, що зв’язані з операціями сігнатурі за правилами аксіономатики:

$$\mathrm{C}_{\mathrm{P\,K}\ \ R} \mapsto \bar{\Omega}(C_{P\,K}), \tag{2.24}$$

де $\bar{\Omega}(C_{P\,K}) \subset \Omega(\mathrm{C}_{\mathrm{P\,K}})$. У результаті роботи конструктора отримується число $\bar{\Omega}(C_{P\,K}) \in [0;1]$, що відображає ступінь подібності текста.

Висновки по другому розділу

У розділі були розглянуті та адаптовані для роботи з українськими природньомовними текстами низка методів.

Було розроблено конструктивно-продукційну модель, що відображає індивідуальний стиль мовлення автора для подальшого її використання для встановлення авторства текстів.

За матеріалами розділу опубліковано роботи [107, 108, 109, 110, 111, 112, 113].

РОЗДІЛ 3
ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ЗАПРОПОНОВАНИХ МЕТОДІВ ВИЗНАЧЕННЯ АВТОРСТВУ ПРИРОДНЬОМОВНИХ ТЕКСТІВ

3.1 Експеримент з визначення авторства природньомовних текстів методами адаптованого рекурентного аналізу

Мета експерименту. Визначення авторства природньомовних текстів методами адаптованого рекурентного аналізу, за методами та моделлю, представленими у підрозділі 2.1.

Експериментальна база. Для проведення експерименту для навчальної вибірки відібрано твори художньої літератури. Це обґрунтовано чітким уявленням авторського стилю та його індивідуальністю, а також достовірною інформацією про авторство. Для першого експерименту до навчальної вибірки було відібрано 20 творів художніх текстів 10 українських авторів. Контрольна вибірка складається з 3 робіт кожного автора.

Автори: IB – І. Багряний, AV – О. Вишня, MV – М. Вовчок, AD – О. Довженко, HK – Г. Квітка-Основ'яненко, PM – П. Мирний, VN – В. Нестайко, VP – В. Підмогильний, IF – І. Франко, MK – М.Хвильовий.

Виконання експерименту. Наведемо реалізацію модифікованого методу рекурентного аналізу на прикладі «Заповіту» Т. Шевченка (рис. 3.1- рис. 3.4).

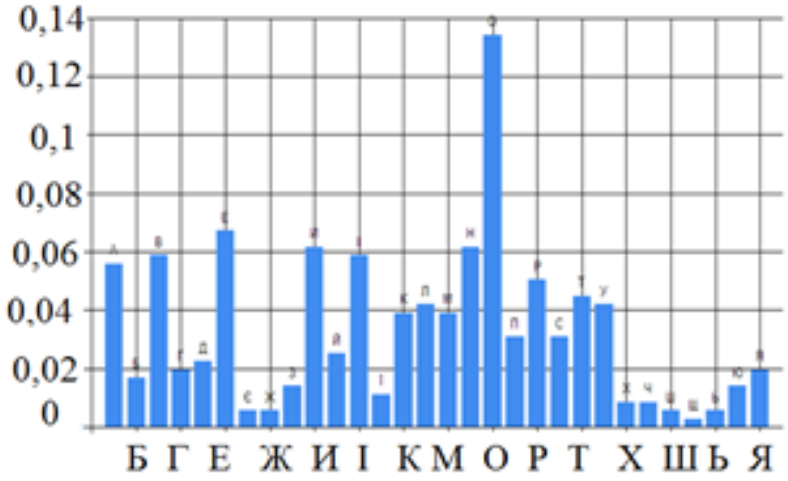

Рисунок 3.1– Діаграма з частотою символів

Обчислення частоти входження кожного символу українського алфавіту наведені на рис. 3.1 у вигляді стовпчикової діаграми.

На рис. 3.2 представлено часовий ряд, сформований на основі обраного тексту з відповідними (як на рис. 3.1) частотами.

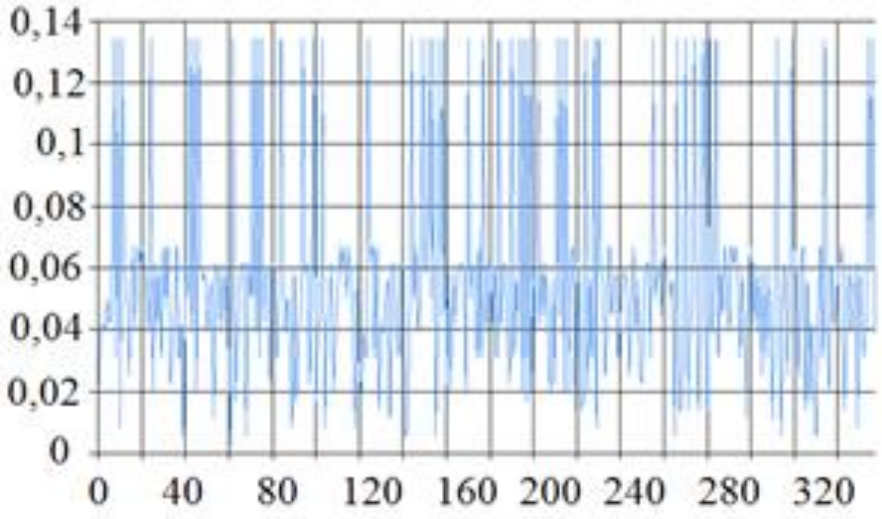

Рисунок 3.2– Часовий ряд тексту «Заповіт»

За отриманими частотами згідно всього тексту «Заповіту» за канонами рекурентного аналізу [148] визначено фазовий простір (рис. 3.3) розмірністю – 2.

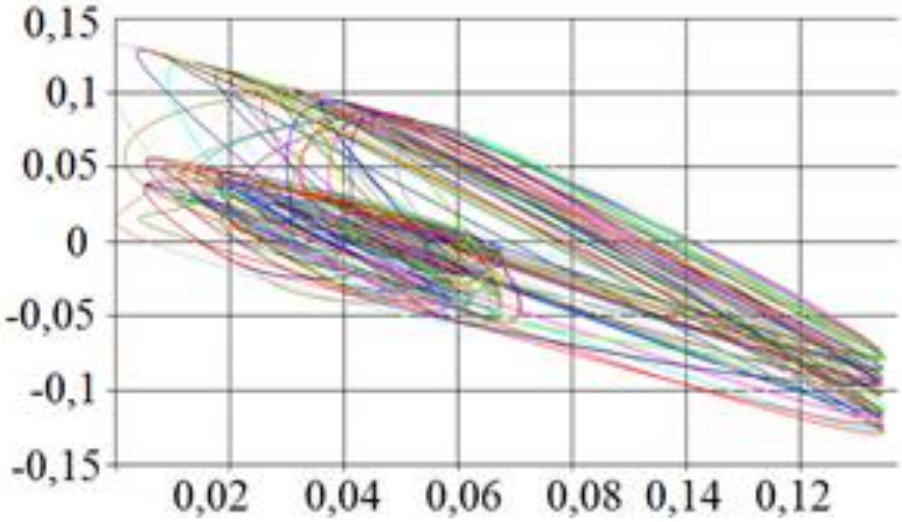

Рисунок 3.3 – Фазовий простір тексту

Побудована рекурентна діаграма має відображати особливості авторського тексту. Діаграма згідно «Заповіту» наведена на рис. 3.4. Значення радіусу околиці точок у фазовому просторі $\varepsilon = 0{,}5$.

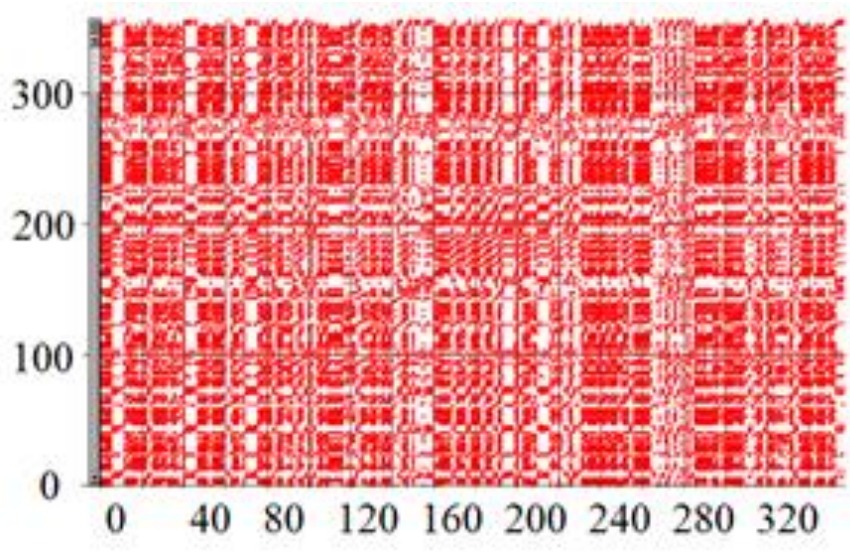

Рисунок 3.4 – Рекурентна діаграма тексту

Для спрощення аналізу діаграми обраховуються показники рекурентності (1)..(8). Для «Заповіту» отримані наступні значення показників (табл. 3.1).

Авторство тексту визначається за найменшою відстанню до еталону, у якості якого прийнято середнє значення за творами автора з навчальної вибірки.

Таблиця 3.1 – Показники рекурентного аналізу «Заповіту»

Назва показників	*Значення*
Міра рекуррентності RR	0,021
Міра детермінізму DET	0,002
Дивергенція DIV	0,125
Средня довжина діагоналей L	2,38
Міра ентропії ENTR	0,769
Міра завмирання LAM	0,00018
Міра затримки TT	2

Вважаємо, що образ X_{ij} належить до класу ω_k, якщо найближчий до X_{ij} образ навчальної вибірки належить ω_i (X_{ij} – вектори у Евклідовому просторі, де i – показник, за яким визначається авторство, j – номер твору в навчальній або контрольній вибірці, x_{ijk} – k-ий елемент вектору X_{ij}).

Класифікація виконується окремо за частотою літер (X_{1j}), довжиною слів (X_{2j}), показниками рекурентного аналізу (X_{3j}), та усіма показниками разом (X_{4j}).

Елементи векторів x_{1jk} – *k*-ий показник рекурентного аналізу (табл. 1); x_{2jk} – частота *k*-ї літери у тексті; x_{3jk} – кількість слів довжиною *k*-літер.

Згідно тексту «Заповіту» отримані значення векторів X_{1j}, X_{2j}, X_{3j}, X_{4j}.

$X_{1,1} = [0.02 \quad 0 \quad 0.13 \quad 2.38 \quad 0.77 \quad 0 \quad 2]$; $X_{2,1} = [0.06 \quad 0.02 \quad 0.06 \quad 0.02 \quad ... \quad 0.02]$;

$X_{3,1} = [14 \quad 11 \quad 2 \quad 18 \quad 11 \quad 11 \quad 7 \quad 4 \quad 4 \quad 1]$;

$X_{4,1} = [0.02 \quad 0 \quad 0.13 \quad 2.38 \quad 0.77 \quad 0 \quad 2 \quad 0.06 \quad 0.02 \ldots \quad 0.02 \quad 14 \quad 11 \quad 2 \quad 18 \quad 11 \quad 11 \quad 7 \quad 4 \quad 4 \quad 1]$.

Для коректності порівняння вектори були унормовані наступним чином:

$$x^*_{ijk} = \frac{x_{ijk} - \min_j(x_{ijk})}{\max_j(x_{ijk}) - \min_j(x_{ijk})}. \qquad (3.1)$$

Отримані результати. У результаті обробки контрольної вибірки були отримані результати (табл. 3.2), де сірим виділені ті результати, що виявили автора твору, або були близькі до нього.

Авторство творів у таблиці пронумеровано наступним чином: 1 – О. Довженко, 2 – І. Багряний, 3 – І. Франко, 4 – М. Коцюбинський, 5 – Л. Українка, 6 – Н. Хвильовий, 7 – О. Вишня, 8 – П. Мирний, 9 – В. Підмогильний, 10 – С. Жадан, 11 – Т. Шевченко.

Інші стовпчики у табл. 2: ЧЛ (частота літер – за вектором X_{2j}); ЛС (кількість літер у слові – за X_{3j}); РА (рекурентний аналіз – за X_{1j}); загальне – результати порівняння за об'єднаним вектором X_{4j}.

У комірках таблиці – інформація щодо визначення найближчих трьох авторів для обраного твору. Якщо перший результат є точними, то наступні не приводяться. Четверте значення визначає близькість першого отриманого результату до реального авторства наступним чином:

$$p = \frac{|l_2 - l_1|}{\max(l_1, l_2)}, \qquad (3.2)$$

де l_1 – відстань між векторами твору та найближчим еталоном, l_2 – відстань між векторами твору та еталоном творів реального автора.

Таблиця 3.2 – Визначення авторства текстів з використанням аналізу за одним символом

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*
2	2	9,2,6/5	6,2,7/15	2
2	8,1,4/10	11,10,5/75	1,2,7/31	10,11,7/27
2	6,9,4/16	9,8,6/36	1,6,9/24	9,8,6/29
7	6,8,2/14	6,9,3/21	8,4,3/9	6,8,9/17
7	2,6,4/13	2,1,7/20	6,7,9/2	2,1,7/16
7	8,7,4/4	7	4,3,9/23	7
1	8,3,9/14	11,7,10/42	8,4,11/39	7,4,11/13
1	1	6,9,8/50	6,2,1/15	6,9,8/32
1	3,7,8/17	8,3,6/17	1	3,8,9/16
10	2,6,10/9	9,6,8/54	1,2,9/33	9,6,2/49
10	10	10	10	10
10	1,9,2/19	10	10	1,10,4/3
4	4	6,9,8/20	8,6,4/36	6,9,8/19
4	1,4,2/5	6,9,3/30	6,9,1/30	6,8,9/20
4	9,1,4/1	8,9,6/22	7,3,6/50	9,8,4/4
5	6,1,4/17	1,4,8/70	8,4,11/19	4,8,1/40
5	4,7,5/7	4,3,7/74	6,7,2/60	4,3,7/48
5	5	11,4,10/65	10,7,1/47	11,5,10/1
8	4,8,3/15	10,11,7/61	8	10,11,7/27
8	8	8	10,1,9/52	8
8	8	8	2,1,9/29	8
9	6,9,3/1	5,10,11/21	10,1,2/32	10,5,11/55
9	1,4,9/8	1,2,7/40	10,1,9/18	1,2,4/30
9	9	1,2,4/39	4,9,1/22	1,2,4/26
3	2,4,9/10	2,1,6/33	1,7,9/15	2,1,9/43
3	5,4,1/8	11,5,10/80	10,1,2/57	5,10,11/40
3	1,9,4/24	8,9,6/25	2,1,9/22	9,6,5/10
6	4,6,1/4	4,3,1/48	6	4,1,3/23
6	6	7,1,4/56	6	7,1,4/31
6	6	9,6,8/8	7,1,2/14	9,6,2/7
11	10,7,1/18	9,6,8/51	6,3,9/78	6,9,8/42
11	11	8,3,4/73	2,7,9/10	8,3,4/42
11	11	11	11	11

Авторство творів у табл. 3.3-табл. 3.4пронумеровано наступним чином: 1 – І. Багряний, 2 – О. Вишня, 3 – М. Вовчок, 4 – О. Довженко, 5 –

М. Коцюбинський, 6 – Г. Квітка-Основ'яненко, 7 – П. Мирний, 8 – В. Нестайко, 9 – В. Підмогильний, 10 – І. Франко, 11 – М. Хвильовий.

Також було виконано визначення автора тексту з використанням N-грамів. Цей метод заснований на розбиття усього тексту на пари сусідніх символів та визначенні їх частоти, з якою вони зустрічаються у творі. При цьому до пари входять символи з нахлестом, тобто спочатку обираються перший та другий символи, потім другий та третій і т.д. Якщо у слові залишається лише один символ, то в пару до нього йде перший символ наступного слова.

Найкращий результат був отриманий при застосуванні 4-грамів (табл. 3.3).

Таблиця 3.3 – Визначення авторства текстів за 4-грамами

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*	*Автор*	*ЧЛ*	*ЛС*	*РА*	*Загальне*
1	1	8	2	1	*6*	6	6	9	6
1	1	7	5	1	*7*	4	9	2	4
1	1	9	9	1	*7*	7	9	3	7
2	2	4	6	2	*7*	7	7	7	7
2	2	8	9	2	*8*	8	9	11	8
2	2	2	5	2	*8*	8	5	4	8
3	3	3	2	3	*8*	8	8	11	8
3	3	7	6	3	*9*	2	2	5	2
3	3	3	3	3	*9*	9	1	5	9
4	4	10	5	4	*9*	9	1	5	9
4	4	7	8	4	*10*	1	1	2	1
4	4	6	8	4	*10*	5	10	5	5
5	5	5	3	5	*10*	10	5	9	10
5	5	5	11	5	*11*	11	7	2	11
5	5	7	3	5	*11*	11	4	2	11
6	6	6	7	6	*11*	9	1	3	9
6	6	6	3	6					

Були проведені експерименти для 2- ... 7-грамів з заміною поетичних творів на прозові. Аналіз даних у табл. 3.3 щодо встановлення авторства за допомогою 4-грамів виявив суттєве покращення аналізу з використанням частоти символів, але зменшення ефективності використання рекурентного аналізу.

Також було виконане порівняння за частотою слів з урахуванням їх закінчень.

Другий стовпчик табл. 4 – ЧС (результати порівняння за вектором X_{1j} *з* даними частоти слів у тексті).

Для виявлення авторства розрахована частота усіх слів у тексті з подальшим формуванням часового ряду, фазового простору та рекурентної діаграми за отриманими даними (табл. 3.4).

Таблиця 3.4 – Визначення авторства текстів за словами

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*	*Автор*	*ЧЛ*	*ЛС*	*РА*	*Загальне*
1	1	8	1	1	*6*	6	6	2	6
1	2	7	9	2	*7*	2	9	3	2
1	1	9	10	1	*7*	7	9	2	7
2	2	4	11	2	*7*	7	7	7	7
2	2	8	6	2	*8*	8	9	11	8
2	2	2	6	2	*8*	2	5	2	2
3	3	3	8	3	*8*	8	8	2	8
3	3	7	2	3	*9*	2	5	1	2
3	3	3	3	3	*9*	9	1	10	9
4	4	10	6	4	*9*	9	1	1	9
4	4	7	1	4	*10*	2	1	1	2
4	2	6	7	2	*10*	2	10	7	2
5	5	5	1	5	*10*	10	5	10	10
5	5	5	10	5	*11*	11	7	2	11
5	5	7	8	5	*11*	2	4	6	2
6	6	6	7	6	*11*	11	1	3	11
6	6	6	11	6					

Дані табл. 3.4 дозволяють стверджувати, що встановлення авторства твору з використанням частоти слів дещо гірше за ефективність аналізу по 4-грамам.

При визначенні авторства текстів контрольної вибірки при першому проведенні експерименту безпомилково визначилися лише автори 12 текстів. Кращий результат визначення авторства дав метод з використанням частоти букв – 12 збігів по автору. Решта методів визначили автора всього у 6 випадках та у 7 по даним рекурентного аналізу.

Відсоток близькості знаходиться у широкому діапазоні від 1% до 80%. Окремо за методами: за даними про частоту літер – 24%, для даних щодо кількості літер у словах – 80%, для рекурентного аналізу тексту – 78% та по результатам порівняння з використанням усіх отриманих даних – 55%.

Також у 22 випадках аналізу тексту автор визначався другим або третім за відстанню. Найкращий показник також за даними щодо частоти літер у тексті, а наступний – за показниками рекурентного аналізу.

Найкращі результати були отриманні при визначенні авторства творів за допомогою 4-грамів та по словах – 85 % та 76 % відповідно за загальним вектором.

3.2 Експеримент зі встановлення авторства природньомовних текстів за кількома класами показників з налаштуванням вагових коефіцієнтів

Мета експерименту. Встановлення авторства природньомовних текстів за кількома класами показників з налаштуванням вагових коефіцієнтів, за методами, представленими у підрозділі 2.4.

Експериментальна база. Розміри навчальної та контрольної вибірок відповідають попередньому експерименту. Умовні позначення: 1 – І. Багряний, 2 – А. Вишня, 3 – М. Вовчок, 4 – А. Довженко, 5 – М. Коцюбинський, 6 – Х. Квітка-Основ’яненко, 7 – П. Мирний, 8 – В. Нестайко, 9 – В. Підмогильний, 10 – І. Франко, 11 – М. Хвильовий.

Виконання експерименту. Експеримент проводився при використанні посимвольного аналізу та із застосуванням аналізу за 4-грамами.

У раніше проведених дослідженнях, найкращий результат визначення авторства тексту давало використання 4-грам. У цьому роботі, порівняння ефективності методів виконувався частотний аналіз 1- і 4-грамм.

Показники в, що формується нами, образ тексту: частота кожного символу в тексті у разі 1-грам (літер) і частота 20% найчастіше зустрічаються 4-грам.

Наведемо приклад формування образу тексту, використовуючи твір Шевченка «Саул», робота містить 2148 аналізованих символів. Результати аналізу тексту представлені як діаграми (рис. 3.5).

Дані розташовані в алфавітному порядку і дозволяють наочно оцінити частоту входження кожної літери, характерну для тексту конкретного автора.

Для оцінки унікальності отриманих значень та їх можливості застосування для визначення авторства слід порівняти отримані дані з інформацією про середню частоту літер українського алфавіту. Згідно з дослідженням [13] літерами українського алфавіту, що найчастіше зустрічаються, є літери О, А, Н. Букви І, Т, В, Е, Р, І, С, К, М матимуть середню частоту поширення.

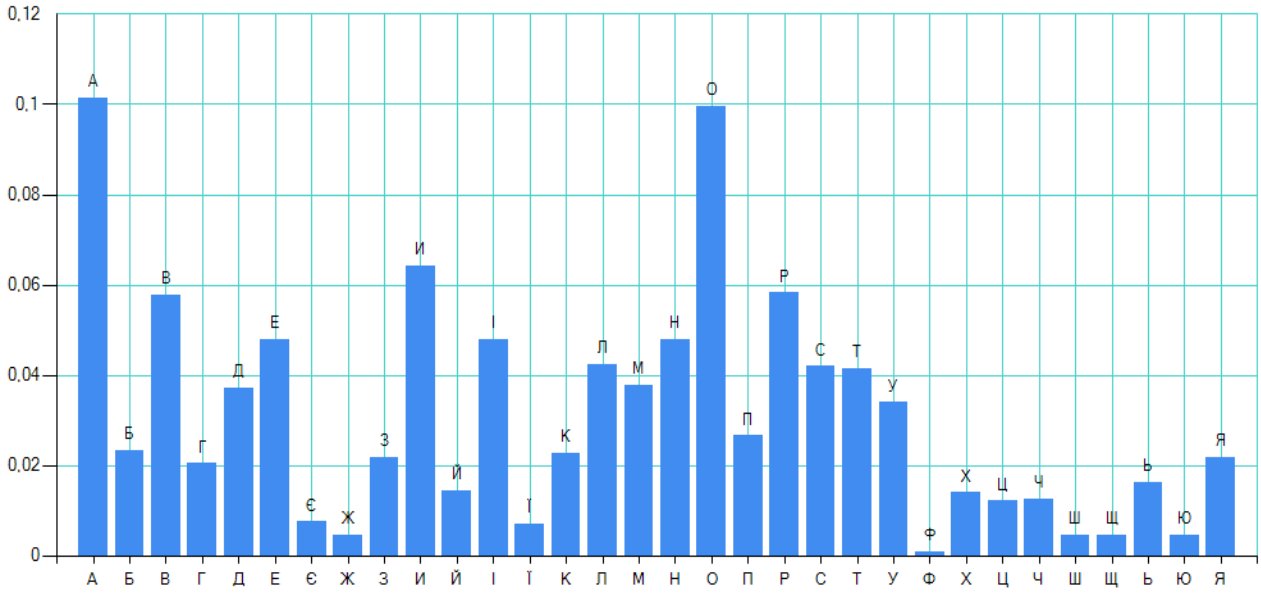

Рисунок 3.5– Діаграма частоти входження символів

Для досліджуваного тексту Т. Шевченка «Саул» характерна найбільша частота входження наступних літер А, О, І, Р, В, Н, І, Е, Л, С. Як можна помітити, існує певна розбіжність із середніми частотами літер у мові та досліджуваному творі, що дозволяє взяти їх як одну із характеристик авторського стилю.

Початкові значення вектору образу тексту Т. Шевченка «Саул», що включають частотні показники: $X' = [0.1014,\ 0.0233,\ 0.0577,\ 0.0205,\ 0.0372,\ 0.0480,\ 0.0074,\ 0.00466,\ 0.0219,\ 0.0642,\ 0.0144,\ 0.0480,\ 0.0070,\ 0.0228,\ 0.0424,\ 0.0377,\ 0.0480,\ 0.0996,\ 0.0265,\ 0.0582,\ 0.0419,\ 0.0414,\ 0.0340,\ 0.0009,\ 0.0140,\ 0.0121,\ 0.0126,\ 0.0046,\ 0.0046,\ 0.0163,\ 0,0047,\ 0.0219,\ \ldots]$.

Наступною характеристикою авторського тексту служить його структурна складність та складність сприйняття. З використанням цього методу аналізу тексту заголовки, підзаголовки і формули найчастіше ігноруються, оскільки не є повноцінними реченнями.

Наведемо приклад значень даних показників щодо вірша Т. Шевченка «Саул» (табл. 3.5).

Таблиця 3.5 – Показники складності сприйняття тексту твору Т. Шевченка «Саул»

Показник	*Значення*
Кількість слів	458
Кількість складів	937
Кількість речень	36
Кількість символів	2148
Ср. кількість слів у реченнях	12.72
Ср. кількість складів у реченнях	26.03
Ср. кількість літер у реченнях	59.67
Ср. кількість складів у словах	2.25
Ср. кількість літер у словах	4.69

Вектор образу тексту Т.Шевченка «Саул» доповнено значеннями показників складності сприйняття: $X' =$[… 12.72, 26.03, 59.67, 2.25, 4.69, …].

До показників складності тексту також було віднесено дані про кількість у тексті слів різної довжини. Для досліджуваного вірша ці показники матимуть такий вигляд табл. 3.6:

Таблиця 3.6 – Дані про кількість слів різної довжини у тексті

Довжина	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Кількість	71	55	30	68	61	60	44	28	15	14	4	4	2	0	1

У досліджуваному тексті найбільше слово складається з 15 літер, слів із 14 літер у тексті немає (табл. 3.6). На основі аналізу всіх наявних текстів було прийнято рішення максимальний довжиною слова вибрати 22, оскільки саме слово такої максимальної довжини є в базі.

Вектор образу тексту Т. Шевченка «Саул» також був доповнений представленими значеннями: $X^{'} =$[… 71, 55, 30, 68, 61, 60, 44, 28, 15, 14, 4, 4, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0,…].

Як і в минулому експерименті, на основі частоти входження літер чи їх послідовностей до тексту формувався частотний ряд, фазовий простір та рекурентна діаграма. Для тексту Т. Шевченка «Саул» вона має наступний вигляд (рис. 3.6).

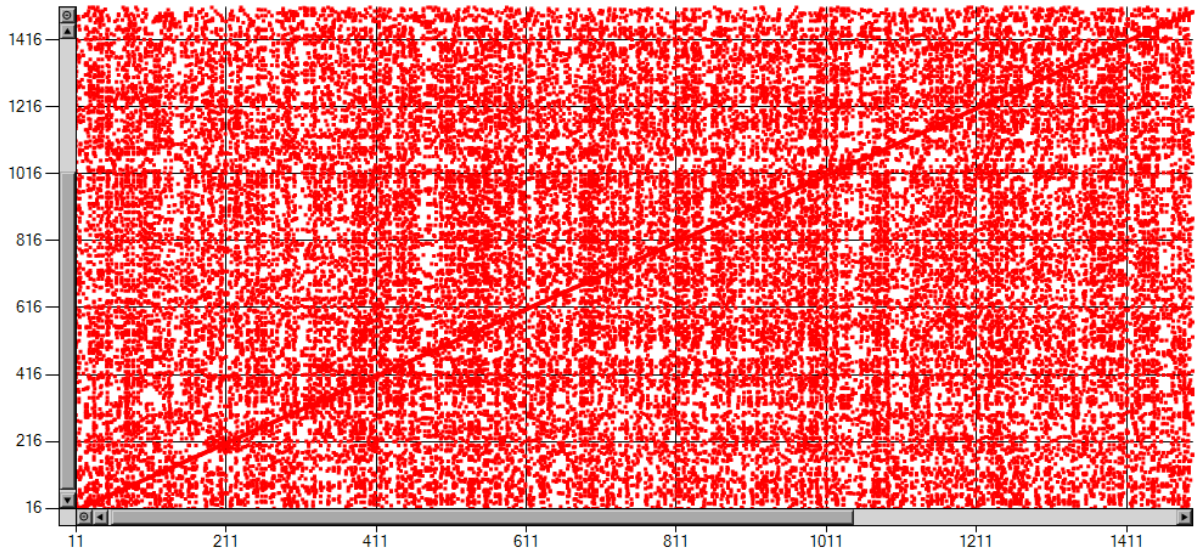

Рисунок 3.6 – Рекурентна діаграма, побудована за текстом Т. Шевченка «Саул»

На основі наведеної діаграми (рис. 3.6) розраховуються показники рекурентності згідно формулам, наведеним вище. Для тексту Т. Шевченка «Саул» було отримано такі значення табл. 3.7.

Вектор образу тексту Т. Шевченка «Саул» доповнено значеннями показників рекурентного аналізу: $X' = [\ldots$ 0,019, 0.002, 0.111, 2.236, 0.6, 8.2, 2.275].

Таблиця 3.7 – Показники рекурентного аналізу для тексту Т. Шевченка «Саул»

Показник	*Значення*
RR	0.019
DET	0.002
DIV	0.111
L	2.236
ENTR	0.6
LAM	8.2
TT	2.275

В образі тексту в експериментах з 1-грамами позиції з 1 по 32 займають значення частот букв, розташовані в алфавітному порядку (рис. 3.6), місця з 33 по 37 займають дані про складність тексту (табл. 3.5), з 38 по 59 – дані про кількість слів різної довжини в тексті (табл. 3.6), з 60 по 66 – дані рекурентного аналізу (табл. 3.7). Таким чином тексту твору Т. Шевченка «Саул», поставлений у відповідність образ-вектор:

$X' =$[0.1014, 0.0233, 0.0577, 0.0205, 0.0372, 0.0480, 0.0074, 0.00466, 0.0219, 0.0642, 0.0144, 0.0480, 0.0070, 0.0228, 0.0424, 0.0377, 0.0480, 0.0996, 0.0265, 0.0582, 0.0419, 0.0414, 0.0340, 0.0009, 0.0140, 0.0121, 0.0126, 0.0046, 0.0046, 0.0163, 0,0047, 0.0219, 12.72, 26.03, 59.67, 2.25, 4.69, 71, 55, 30, 68, 61, 60, 44, 28, 15, 14, 4, 4, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,019, 0.002, 0.111, 2.236, 0.6, 8.2, 2.275].

Отримані результати. Результати експерименту з 1-грамами представлені у табл. 3.8. Затемнені осередки – випадки, у яких автор твори з контрольної вибірки було визначено правильно.

Працюючи з 1-грамами сформовано 15 поколінь хромосом. Обчислені вагові коефіцієнти дозволили покращити результат визначення авторства тексту вчетверо з 6 до 24 з 33 творів всього, що становило покращення з 18% до 80%.

У ході експерименту було отримано 64 хромосоми, які давали найбільшу кількість чітко визначених авторів. Для отримання більш повної картини значущості досліджуваних показників далі наведені дані про інтервали вагових коефіцієнтів для кожного з показників та їх середні значення (табл. 3.8-табл 3.11).

Таблиця 3.8 – Результати визначення авторства текстів з використанням вагових коефіцієнтів (ВК) та без них посимвольно

Автор твору		*0*	*1*	*1*	*2*	*2*	*2*	*3*	*3*	*3*	*4*	*4*	*4*	*5*	*5*	*5*	*6*	*6*	*6*
Автор, що визначен	*без ВК*	1	9	9	2	2	3	9	10	11	3	11	5	7	3	3	10	9	11
	з ВК	1	1	1	2	9	2	3	10	3	4	4	4	4	5	5	6	5	5

Автор твору		*7*	*7*	*7*	*8*	*8*	*8*	*9*	*9*	*9*	*10*	*10*	*10*	*11*	*11*	*11*
Автор, що визначен	*без ВК*	7	5	7	6	3	10	7	3	11	8	9	10	10	9	11
	з ВК	7	11	7	8	11	8	9	9	9	10	10	10	11	5	11

Згідно з даними, представленими у табл. 8 серед вагових коефіцієнтів всіх хромосом у категорії частоти символів найбільше значення отримали літери Ш, Ф, Ї, Я, Є.

Варто зазначити, що дані літери не є найчастішими літерами українського алфавіту, що дозволяє вважати їхню частоту інформативною характеристикою конкретного тексту та авторського стилю в цілому. Вагові коефіцієнти даних показників, крім частоти літери Я, також демонструють значні коливання значеннях, що лише підтверджує раніше зроблений висновок.

Таблиця 3.9 – Розрахункові вагові коефіцієнти для частот букв українського алфавіту

Літера	*Макс*	*Мін*	*Ср*	*Літера*	*Макс*	*Мін*	*Ср*
А	1,726	1,399	1,596	Н	2,442	2,175	2,332
Б	2,550	2,299	2,426	О	3,842	2,309	3,261
В	1,140	0,921	1,060	П	3,590	3,270	3,432
Г	3,287	3,002	3,135	Р	4,039	3,679	3,848
Ґ	1,851	1,737	1,800	С	2,675	2,518	2,592
Д	0,759	0,699	0,729	Т	0,984	0,000	0,692
Е	0,000	0,000	0,000	У	1,221	1,088	1,160
Є	5,049	4,599	4,819	Ф	5,509	5,165	5,385
Ж	2,643	2,493	2,564	Х	2,032	1,517	1,781
З	4,842	2,830	3,728	Ц	3,313	3,070	3,180
И	4,076	3,882	3,962	Ч	4,879	4,497	4,652
І	1,331	0,392	0,894	Ш	5,544	5,110	5,327
Ї	5,171	4,812	4,988	Щ	3,850	3,549	3,705
Й	3,881	3,545	3,703	Ь	3,850	3,567	3,738
К	1,148	1,114	1,128	Ю	1,233	0,866	1,009
Л	0,215	0,203	0,208	Я	5,088	0,805	1,587
М	1,072	0,988	1,030				

Частоти ж букв Е, Л, Д, Т, М – букви, частоти яких мають найменші вагові коефіцієнти серед усіх. А літери Я, З, О, Т, І мають найбільший розкид у значеннях вагових коефіцієнтів за весь час експерименту.

Згідно табл. 3.9 найбільші вагові коефіцієнти мають слова довжиною 1, 3, 4, 5 та 15, 19, що може вказувати на службові слова, що зазвичай мають невелику довжину, та занадто довгі слова, що можуть вказувати на особливість авторського стилю мовлення.

Таблиця 3.10 – Розрахункові вагові коефіцієнти для слів різної довжини

Довжина	*Макс*	*Мін*	*Ср*	*Літера*	*Макс*	*Мін*	*Ср*
1	5,156	4,696	4,929	12	2,835	1,350	2,336
2	1,010	0,900	0,953	13	4,920	2,218	3,871
3	5,204	2,882	4,479	14	4,599	1,893	3,522
4	4,920	4,679	4,781	15	4,574	4,342	4,458
5	5,529	5,096	5,299	16	0,958	0,912	0,930
6	1,322	1,230	1,278	17	3,136	2,856	2,976
7	1,807	0,406	0,795	18	0,750	0,685	0,717
8	2,378	2,270	2,338	19	4,323	4,055	4,201
9	3,702	3,260	3,467	20	4,171	3,864	4,037
10	3,287	2,928	3,139	21	4,012	3,654	3,817
11	1,552	1,431	1,489	22	3,802	3,567	3,702

Згідно з даними табл. 3.10, найбільший ваговий коефіцієнт має показники TT – показник затримки, що відображає середню довжину повторень поспіль послідовностей статистично близьких символів. Однак, ваговий коефіцієнт цього показника також продемонструвала значне коливання.

Таблиця 3.11 – Розрахункові вагові коефіцієнти для рекурентних показників

Покажчик	*Макс*	*Мін*	*Ср*
DET	3,458	1,053	2,124
DIV	4,394	3,914	4,167
ENTR	4,166	3,728	3,990
L	3,874	3,602	3,714
LAM	4,082	1,749	3,361
RR	1,684	1,566	1,628
TT	4,651	2,140	4,086

Другим за ваговими коефіцієнтами є показник DIV (дивергенція) – величина, зворотна L максимальній довжині діагональних структур, що в рамках тексту відображає кількість послідовностей символів, що повторюються поспіль,

в даному тексті. Ваговий коефіцієнт цього показника досить стабільна, що дозволяє говорити про цей показник, як про важливий параметр тексту.

Показники ENTR (ентропія) та L (середня довжина діагональних ліній) мають менший ваговий коефіцієнт, проте його коливання також незначне.

Показник завмирання LAM згідно з отриманими даними також має значний ваговий коефіцієнт, але через його сильне коливання, показник не може бути достовірною характеристикою тексту.

Аналізуючи отримані дані (табл. 3.8-табл. 3.11) найбільші серед усіх вагові коефіцієнти в категорії складності тексту мають показники кількості слів завдовжки 5, 3 та 1, а також кількість складів у словах та літер у реченнях. З іншого боку, вагові коефіцієнти даних показників немає значних коливань протягом усього експерименту.

Однак для даних про довжину слів, слова довжиною 3 спостерігається значне розходження, тобто інтервал між максимальним та мінімальним вагових коефіцієнтів даного показника для різних хромосом.

Таблиця 3.12 – Розбіжності вагові коефіцієнти для показників складності тексту

Покажчик	*Макс*	*Мін*	*Ср*
слова у реченні	4,721	4,390	4,536
літери у словах	4,309	3,041	3,446
слова у реченнях	0,319	0,297	0,306
склади у реченнях	0,862	0,793	0,829
склади у словах	4,848	4,509	4,670

Останній показник у хромосомі – вагові коефіцієнти для ε – під час проведення експерименту значних коливань не продемонстрував. Максимальним значенням вагового коефіцієнту стало 2,842, мінімальним – 0,053. Середнє значення для вагового коефіцієнту ε стало 1,095. На основі цих даних можна зробити висновок про те, що значне коливання ε призведе до того, що близькі за частотою послідовності символів стануть невиразними. Як наслідок, рекурентна діаграма буде спотворена і не зможе повною мірою відобразити авторський стиль.

Також був проведений експеримент при використанні 4-грам, як найрезультативнішого варіанту в попередньому експерименті. В результаті було отримано такі дані (табл. 3.13). Затемнені осередки – випадки, у яких автор твори з контрольної вибірки було визначено правильно.

Працюючи з аналізом тексту з допомогою 4-грам під час експерименту для отримання результату було сформовано 28 поколінь. Отримані вагові коефіцієнти дозволили поліпшити результат визначення авторства тексту лише трохи з 27 до 30 з 33 творів всього, що становило поліпшення з 82% до 91%.

Таблиця 3.13 – Результати визначення авторства текстів на основі 4-грам

Автор твору		1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
Автор, що визначен	*без ВК*	0	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	5
	з ВК	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	5
Автор твору		7	7	7	8	8	8	9	9	9	10	10	10	11	11	11			
Автор, що визначен	*без ВК*	7	7	4	8	9	8	9	9	9	10	1	1	11	11	9			
	з ВК	7	7	7	8	8	8	9	9	9	10	1	1	11	11	11			

При застосуванні генетичного алгоритму для пошуку вагові коефіцієнти до різних груп параметрів результати визначення авторства природньомовних текстів значно покращилися. При аналізі тексту 1-грам - у 4 рази (з 6 до 24 збігів за автором). При роботі з 4-грамами результат також був покращений, але меншою мірою – з 27 до 30 збігів. Отримані результати склали 80% та 91% відповідно.

Застосування зазначеної методики покращило результат визначення авторства в обох випадках, що дозволяє говорити про її ефективність.

Якщо розглядати представлені групи показників та їх вагові коефіцієнти у визначенні авторства тексту, то найбільш важливими, згідно з отриманими ваговими коефіцієнтами, будуть частоти букв Ф, Ш, Ї, Є, Ч та показники складності тексту, а саме кількість у тексті слів довжиною 5, 1 і 4 літери та кількість складів у ловах та літер у реченнях. Для рекурентного аналізу найбільш інформативними стали показники дивергенції, затримки та ентропії.

3.3 Експеримент зі встановлення авторства природньомовних текстів на основі конструктивно-продукційного моделювання

Мета експерименту. Встановлення авторства природньомовних текстів на основі конструктивно-продукційного моделювання, за методами та моделями, представленими у підрозділі 2.5.

Експериментальна база. Розміри навчальної та контрольної вибірок для першої половини експерименту відповідають попередньому експерименту.

Для другої половини експериментів, роботі з граматиками, обидві вибірки були подвоєні, відповідно навчальна вибірка включала 40 робіт тих самих авторів, а 60 текстів склали нову контрольну вибірку – по 6 робіт кожного автора.

Автори: IB – І. Багряний, AV – О. Вишня, MV – М. Вовчок, AD – О. Довженко, HK – Г. Квітка-Основ'яненко, PM – П. Мирний, VN – В. Нестайко, VP – В. Підмогильний, IF – І. Франко, MK – М.Хвильовий.

Виконання експерименту. У цьому експерименті досліджується метод визначення авторства текстів, заснований на структурі речень індивідуальної мови автора.

Використовується стохастична граматика для створення правил, що описують структуру речень тексту. До кожного правила визначається можливість застосування у конкретному творі. Імовірність виведення всього речення визначається як добуток ймовірностей, послідовностей частин мови які є у ньому. Отримані правила породжуватимуть мову, характерну для оброблюваного та структурно подібних творів певного автора.

Для опису структури досліджуваного тексту використовуються частини мови як характеристику слова. Таким чином, кожне слово у реченні замінюється на частину мови, якою він є.

Для тегування слів у тексті українською мовою використовувалися наступні теги: дієслово (гол), іменник (сущ), займенник (місць), прикметник (прил), спілка, прислівник (нар), прийменник (предл), причастя (прич), вигук , частка, дієприслівник (дієприч).

Для кожної частини мови прораховується ймовірність її появи у певному місці речення у цьому тексті. Імовірність появи певної частини мови в досліджуваній послідовності дозволить більш точно вловити індивідуальний стиль листа, характерний кожному з авторів, що досліджуються. Після отримання тексту у вигляді набору послідовностей частин мови в реченнях з ймовірністю їх появи в конкретному місці формуються правила.

Для цього всі речення, що починаються на ту саму частину мови, групуються, перше слово відкидається, і процедура прорахунку ймовірності повторюється для наступного слова.

Після речення знову групуються згідно частин мови на початку, перше слово знову відкидається і вважається ймовірність для наступного елемента і т.д. Імовірність прораховується як кількість наявних випадків у тексті, поділена на їхню загальну кількість.

Таким чином, правила підстановки для деякого твору Т мають початковий нетермінал, далі термінали, що відповідають кожному слову у реченні та ймовірність застосування відповідного правила при розборі твору та мають вигляд:

$$\sigma \xrightarrow{p_{1j}} b_{1j} A_{1,j}, \; A_{i,j} \xrightarrow{p_{i+1,k}} b_{i+1,k} A_{i+1,k}, \; j = 1 \ldots J_i, \; k = 1 \ldots K_i$$

де σ – початковий нетермінал, b_{ij} – термінали, відповідні *i*-му слову у реченні (і відповідні *i*-му правилу, що застосовується при розборі речення або *i*-му рівню правила), $A_{i,j}$ – *j*-й нетермінал у правилі *i*-го рівня, p_{ik} – ймовірність застосування відповідного правила при розборі даного твору, J_i, K_t – кількість різних нетерміналів у правій частині правил *i-1*-го рівня та *i*-го рівня, відповідно.

Рівень відповідає порядковому номеру слова у реченні.

При використанні даного методу тексту речення з твору «Етюд» І. Багряного представлена у вигляді послідовності частин мови, що входять до нього, матиме наступний вигляд табл. 3.14.

Приклад одного з автоматично відновлених правил наведено нижче Правило описує всі 24 речення у творі «Етюд» І. Багряного, що починаються з дієслова. Як видно з представлених ймовірностей, у досліджуваному творі 31% речень починатиметься саме з дієслова. А відсоток речень, що складаються лише з одного слова, дієслова, становить 17%.

Допускається кілька альтернативних правил з нетерміналом у лівій частині правила, але при цьому термінали у правій частині таких правил різні, що забезпечує детермінований розбір.

Таблиця 3.14 – Правила стохастичної граматики щодо «Етюду» І. Багряного

Ліва частина	*Ймовір ність*	*Термінал*	*Нетер мінал*	*Ліва частина*	*Ймовір ність*	*Термінал*	*Нетер мінал*
σ	0,31	дієсл	$A_{1,1}$	$A_{3,4}$	1,00	прийм	$A_{4,3}$
$A_{1,1}$	0,17	ε		$A_{3,5}$	1,00	присл	$A_{3,3}$
$A_{1,1}$	0,13	ім	$A_{2,1}$	$A_{3,6}$	1,00	дієсл	$A_{4,4}$
$A_{1,1}$	0,04	прийм	$A_{2,2}$	$A_{3,7}$	0,80	прикм	$A_{3,7}$
$A_{1,1}$	0,04	союз	$A_{2,3}$	$A_{3,7}$	0,20	ε	
$A_{1,1}$	0,21	дієсл	$A_{2,4}$	$A_{4,1}$	1,00	частка	$A_{5,1}$
$A_{1,1}$	0,13	займ	$A_{2,5}$	$A_{4,2}$	1,00	прикм+ім	G3
$A_{2,1}$	0,40	ε		$A_{4,3}$	1,00	ім	$A_{5,2}$
$A_{2,1}$	0,40	ім	$A_{3,1}$	$A_{4,4}$	1,00	присл	$A_{5,3}$
$A_{2,1}$	0,20	прийм	$A_{3,2}$	$A_{4,5}$	1,00	дієсл	$A_{5,4}$
$A_{2,2}$	1,00	ім	$A_{3,3}$	$A_{5,1}$	1,00	ім	$A_{6,1}$
$A_{2,3}$	1,00	дієсл	$A_{3,1}$	$A_{5,2}$	0,33	ε	
$A_{2,4}$	0,20	прикм+ім	$A_{4,5}$	$A_{5,2}$	0,33	дієсл	$A_{6,2}$
$A_{2,4}$	0,20	ε		$A_{5,2}$	0,33	присл	$A_{6,1}$
$A_{2,4}$	0,20	прийм	$A_{2,1}$	$A_{5,3}$	1,00	ім	$A_{3,3}$
$A_{2,4}$	0,20	дієсл	$A_{3,4}$	$A_{5,4}$	1,00	прийм	$A_{6,3}$
$A_{2,4}$	0,20	займ	$A_{3,5}$	$A_{6,1}$	1,00	дієсл	$A_{7,1}$
$A_{2,5}$	0,33	ім	$A_{3,6}$	$A_{6,2}$	1,00	дієсл	$A_{7,3}$
$A_{2,5}$	0,33	прикм	$A_{3,7}$	$A_{6,3}$	1,00	дієприкм	$A_{7,2}$
$A_{2,5}$	0,33	дієсл	$A_{3,4}$	$A_{7,1}$	1,00	прийм	$A_{5,3}$
$A_{3,1}$	0,50	ε		$A_{7,2}$	0,33	прикм+ім	$A_{7,2}$
$A_{3,1}$	0,20	союз	$A_{4,1}$	$A_{7,2}$	0,67	ε	
$A_{3,2}$	1,00	займ	$A_{4,2}$	$A_{7,3}$	1,00	ім	$A_{3,3}$
$A_{3,3}$	1,00	ε					

Таким чином, текст представляється у вигляді набору правил, що описують його структурні особливості за допомогою описаних вище правил. Символ ε означає пусто (кінець правила).

Приклад перших кількох правил згідно з таблицею мають такий вигляд: $\sigma \xrightarrow{0.31} \text{гл}A_{1,1}; A_{1,1} \xrightarrow{0,17} \varepsilon;\ A_{1,1} \xrightarrow{0,13} \text{сущ}A_{2,1}$. У лівій частині правила нетермінал, далі вказується ймовірність його застосування та у правій частині правила термінал з нетерміналом для переходу до наступного правила.

Для першого експерименту вибірки співпадали по складу з попереднім. Для проведення другого експерименту обидві вибірки були збільшені вдвічі, відповідно до навчальної вибірки увійшли по 40 творів тих самих авторів, і 60 текстів склали нову контрольну вибірку – по 6 назв одного автора.

Таблиця 3.15 – Тегування речень та відповідні ймовірності правил стохастичної граматики

Слово у реченні	*Тег*	*Ймовірність*
Чорні	прикм	0,06
ґрати	ім	0,6
розпанахали	дієсл	0,6
небо	ім	0,125

Отримані результати. Нижче на рис. 3.7 та рис. 3.8 представлені результати кожного експерименту. Кожен стовпчик на діаграмі представляє роботи конкретного автора із контрольної вибірки. Стовпець розбитий на дві зони, де синя частина відображає кількість тестів з чітко визначеним авторством, а помаранчева – з хибним.

Відповідно до отриманих результатів, при роботі з меншою з двох вибірок, що містить у загальній кількості 200 творів 10 авторів у навчальній вибірці (по 20 для кожного автора), та 30 твори у контрольній, кількість випадків правильного встановлення авторства – 24, що становить 80%.

Найкращий результат було отримано під час роботи з творами О.Довженка, П.Мирного, В.Нестайка та В.Підмогильного. Автором із найбільш складним стилем виявився І.Франко.

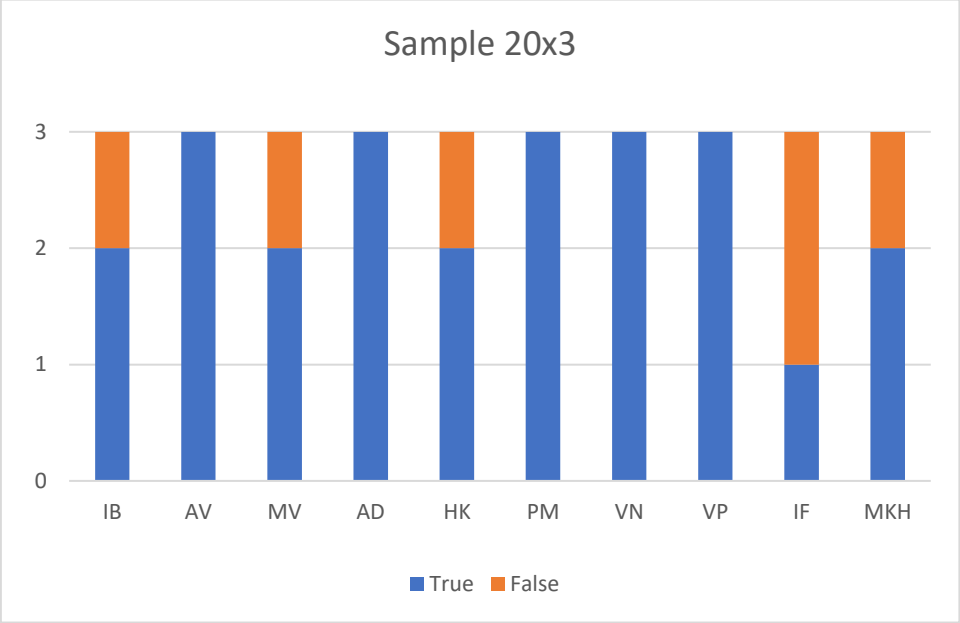

Рисунок 3.7 – Результати встановлення авторства з вибіркою 20x3

Згідно з отриманими результатами, при роботі з більшою вибіркою, збільшеною вдвічі (400 творів 10 авторів у навчальній вибірці, та 60 твори в контрольній), кількість випадків правильного встановлення авторства – 45, що становить 75%.

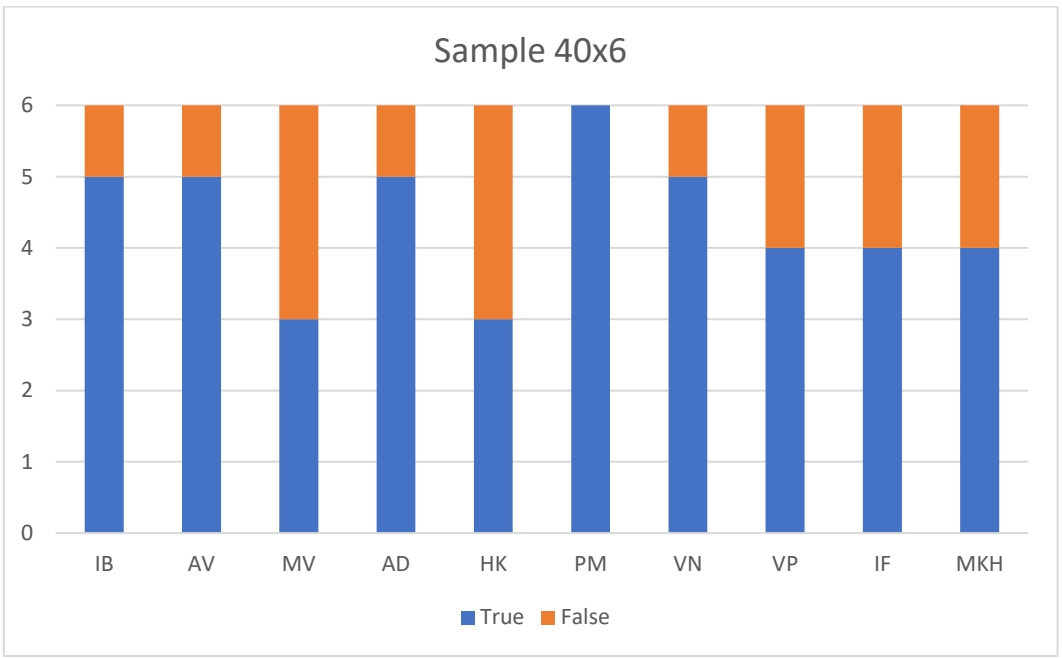

Рисунок 3.8 – Результати встановлення авторства з вибіркою 40x6

Найкращий результат було отримано під час роботи з творами П.Мирного. Авторами з найбільш складним стилем виявилися Г. Квітка-Основ'яненко та

М Вовчок – тільки у половина їх творів з вибірки було правильно визначено авторство. Для більш достовірного результату було введено довірчий інтервал табл. 3.16.

Табл.иця 3.16 – Довірчий інтервал авторів

	IB	*AV*	*MV*	*AD*	*HK*	*PM*	*VN*	*VP*	*IF*	*MKH*
Середнє	0,83	0,83	0,61	0,83	0,58	0,97	0,87	0,69	0,67	0,66
Макс	0,86	1,00	0,67	0,86	0,60	0,99	0,89	0,70	0,71	0,69
Мін	0,80	0,64	0,55	0,80	0,56	0,95	0,85	0,68	0,63	0,63
Діапазон	0,05	0,37	0,12	0,06	0,05	0,04	0,03	0,02	0,07	0,07

Довірчі інтервали різних авторів суттєво відрізняються між собою. Так, в середньому, интревал коливатися від 0,04 до 0,08, проте, у О. Вишня і М. Вовчок він значно більше, і становив 0,37 і 0,12 відповідно. Мінімальний інтервал становив 0,02 у В. Підмогильного.

Для деяких з авторів, таких як Г. Квітка-Основ'яненко та М. Вовчок характерний особливий стиль оповідання, що складно піддається класифікації та структуризації, завдяки чому для них характерний низький результат розпізнавання. Тоді як автори І. Багряний, О. Вишня, О. Довженко, П. Мирний та В. Нестайко мають більш індивідуальний стиль листа, що відображається в структурі речень і дозволяє встановлювати їх авторство з високою точністю.

З урахуванням довірчих інтервалів було отримано результати табл. 3.17.

Таблиця 3.17– Результати визначення авторства з урахуванням довірчого інтервалу

З інтервалом	1 автор, дійсний	1 автор, не дійсний	більше 1 автора, серед них є дійсний	бальше 1 автора, среди них немає дійсного
Кіл-ть	14	2	36	8
Відсоток	23,3%	3,3%	60%	13,3%

Згідно з отриманими результатами, для 14 випадків у результаті визначення авторства було отримано одного вірного кандидата, що склало 23,3%. У 36 випадках програма змогла звузити кількість претендентів до 3, при цьому в 31 випадку претендент із найбільшою структурною схожістю був вірним автором

(що становило 51,7% від загальної вибірки) і ще у 5 випадках вірний автор потрапив до списку кандидатів (8,3 % від загальної вибірки). Також всього у 8 випадках авторство тексту не було визначено – жоден з представлених кандидатів не був вірним, що становило 13,3%.

В результаті аналізу отриманих даних можна стверджувати, що розмір довірчого інтервалу можна також вважати характерною рисою особистого стилю автора. Так, у деяких із них розмір довірчого інтервалу значно відрізняється від інших авторів. У О. Вишня і М. Вовчок він значно більший. Примітно, що для О. Вишня характерний власний стиль листа, що дозволило з достатньою точністю визначати тексти його авторства, тоді як стиль М. Вовчок досить важкий для класифікації. Мінімальний інтервал був отриманий для робіт автора В. Підмогильного, що, однак, призводить до досить високих результатів його класифікації.

Таким чином, відсоток вірного визначення авторства з урахуванням довірчого інтервалу було покращено до 83,33% (у 50 випадках із 60 автор твору було правильно розпізнано).

При аналізі текстів із вибіркою 20х3 кількість збігів за автором становила 24 із 30 робіт. Працюючи зі збільшеними вдвічі вибірками результат також позитивний, але у меншою мірою – 45 з 60 збігів. Отримані результати склали 80% та 75% відповідно.

З урахуванням довірчого інтервалу результати вдалось покращити до 83,3%. В результаті аналізу можна стверджувати, що розмір довірчого інтервалу можна вважати характерною рисою особистого стилю автора. Так, великий розмір довірчого інтервалу може говорити про низький рівень диференціації авторського стилю і, як наслідок, поганий результат щодо авторства його робіт. І навпаки – за маленького довірчого інтервалу ймовірність впевненої диференціації стилю автора значно підвищується. Знаковим також є значення середнього значення схожості творів у навчальній вибірці – що вище значення, то чіткіше визначається стиль автора і, вище результат визначення творів його авторства.

3.4 Експеримент зі встановлення ефективності методів та засобів визначення значими показників профілю автора природномових текстів

Мета експерименту. Встановлення ефективності методів та засобів визначення значими показників профілю автора природномових текстів, за методом та моделями, представленими у підрозділі 2.5.

Експериментальна база. Розміри навчальної та контрольної вибірок співпадають з другою вибіркою (в обсязі 400 творів навчальною та 60 контрольної) експерименту, представленому у підрозділі 3.3.

Виконання експерименту. Наведено методи формування та оптимізації профілів авторів. Профіль автора це образ – вектор у багатовимірному просторі, компоненти якого є вимірами текстів автора рядом методів на основі 4-грам, стемування, рекурентного аналізу та формальної стохастичної граматики. Профіль автора є моделлю його мови, включаючи словниковий запас, особливості синтаксису речень. Проводиться порівняльний аналіз ефективності кожного з методів. Засобами генетичного алгоритму формується усічений профіль автора. Виключаються незначні показники, що дозволяє скоротити їхню кількість на 20%. Усічений профіль автора містить важливу для даного автора атрибутику і є ефективною атрибуцією конкретного автора.

У процесі підбору вагових коефіцієнтів для кожного з показників за допомогою генетичного алгоритму виконуються наступне: випадковим чином формується початковий вектор вагових коефіцієнтів W_k першого покоління, визначається фітнес-функція, відбір кращих з кросовером і мутацією для формування нового покоління W_k.

Фітнес-функція $\sum_{k=1}^{40} \rho(W_k' \cdot X_k)$, де X_k – профіль автора k-того твору, W_k' – відповідні цьому автору вагові коефіцієнти вимірювань, ρ – функція, яка експериментально визначає чи правильно встановлено авторство k-того твору.

Останні два кроки повторюються до тих пір, поки не припиниться покращення результату функції, після чого процес вважається завершеним, а вагові коефіцієнти визначеними.

Останній етап – скорочення кількості показників. Послідовно виключається x_j та w_j такі, що $w_j = \min_k(w_k)$. Якщо результат залишається таким самим або трохи погіршується – скорочення профілю продовжується. Як тільки результат починає погіршується значно – скорочення зупиняється та вважається завершеним.

Приклад 4-грам із твора «Доля» Т.Г. Шевченко:

Ти не лукавила зо мною,

Ти другом, братом і сестрою…

Отримані 4-грами: тине, инел, нелу, елук, лука, укав, кави, авил, вила, илаз, лазо, азом, зомн, омно, мною, ноют, оюти, ютид,…

У цьому експерименті також застосовано та досліджено з точки зору своєї ефективності для визначення авторства адаптований до української мови стемер Портера [6, 48]. Він використовується для роботи безпосередньо з текстами різних авторів і побудовою профілю частоти використання різних основ, характерний для кожного автора.

Приклад тем того ж уривка з твору «Доля» Т.Г. Шевченко: т, лукав, мн, друг, брат, сестра.

На його основі було збудовано комплексний словник, що містить унікальні основи слів, їх закінчення та префікси. Для зменшення його розміру було проведено попередню вибірку унікальних списків закінчень та присвоєння основі слова лише індексу з нього. Введено список чергування голосних у словах.

Для створення списків префіксів для основ проведено аналіз сформованого словника на наявність основ, що відрізняються лише наявністю префіксу простим перебором. В результаті початковий словник основ зменшився – всім ключовим основам присвоєно відповідний індекс зі списку префіксів, а зайві основи з ними видалено.

Перевагою отриманого словника є підтримка обліку всіх словоформ для основ, кожної з яких буде присвоєно унікальний індекс. Таким чином, усі

відмінки, різні форми слів, а також отримані додаванням приставки слова безпомилково будуть вести до єдиної основи.

Використовується стохастична граматика для створення правил, що описують структуру речень тексту.

Для опису структури досліджуваного тексту використовуються частини мови як характеристику слова. Таким чином, кожне слово у реченні замінюється на частину мови, якою він є. Для отримання більшої інформації про структуру речень та правила їх побудов, характерні для певного автора зчитуванні не тільки частини мови, а й форма, число, рід тощо. для досліджуваного слова.

Для кожної частини мови прораховується ймовірність її появи у певному місці речення у цьому тексті. Імовірність появи певної частини мови в досліджуваній послідовності дозволить більш точно вловити індивідуальний стиль мовлення, характерний кожному з авторів, що досліджуються. Після отримання тексту у вигляді набору послідовностей частин мови в реченнях з ймовірністю їх появи в конкретному місці формуються правила.

Приклад правила того ж уривка із твору «Доля» Т.Г. Шевченка:

$$\sigma \xrightarrow{\text{p1}} \text{займ}A_{1,1};\ A_{1,1} \xrightarrow{\text{p2}} \text{ч}A_{2,1};\ A_{2,1} \xrightarrow{\text{p3}} \text{дієсл}A_{3,1};$$

$$A_{3,1} \xrightarrow{\text{p4}} \text{прийм}A_{4,1};\ A_{4,1} \xrightarrow{\text{p5}} \text{займ}A_{5,1};\ldots$$

де σ – початковий нетермінал, $A_{i,j}$ – j -й нетермінал в правилі i -го рівня, p_i – ймовірність застосування відповідного правила при розборі цього твору.

Для отримання профілю конкретного автора проводяться розрахунки визначення кожного з досліджуваних груп показників всім творів автора в навчальної вибірці. Далі вони всі збираються до одного вектора X – це і є профіль автора.

Наприклад, при роботі з 4-грамами, на основі отриманих показників формується вектор, який містить частоти входження кожного такого 4-грама в тексті. Для складання профілю автора, в розрахунок беруться такі вектори для кожного з творів навчальної вибірки і розраховується середнє значення для

кожного з них. Подібна процедура повторюється для формування векторів з урахуванням інших груп показників.

Приклад вектору-образу Т.Г. Шевченка на основі 4-грам:Y' =[АБАЗ, АБАЙ, АБАР, АБАС, АБАТ, АБАУ, АБЕР, АБІК, АБІЛ, АБЛА, АБОГ, АБОТ, АБОЮ, АБОЯ, АБУД, АБУД, АБУЛ, АБУС, АБУТ, АВАБ, АВАВ, АВАЛ, …].

Всього у векторі 8748 4-грам, що використовуються в тексті. І їх частоти:

X' =[0.0001249, 0.0001565, 0.0001249, 0.0001565, 0.0001249, 0.0001249, 0.0001565, 0.0001565, 0.0001249, 0.0001249, 0.0004998, 0.0001249, 0.0001249, 0.0001249, 0.0004381, 0.0001249, 0.0001249, 0.0001249, 0.0001565, 0.0002499, 0.0001565, 0.0004696, …].

Як видно, отриманих 4-грам та їх частот велика кількість, працювати з якою витратно за часом та обчислювальними ресурсами. Однак, оскільки для кожного автора характерний свій стиль написання, для різних авторів можуть бути найбільш інформативні різні 4-грами. Крім цього, буквосполучення, що зустрічаються найчастіше можуть мати найбільше значення, оскільки будуть характерною особливістю авторської мови. Так, перелік отриманих частот вимагає додаткового аналізу їх інформативності та подальшого скорочення даних для роботи лише з найбільш значущими показниками.

Для оптимізації роботи та отримання кращого результату, під час роботи з різними показниками у векторах, був застосований генетичний алгоритм визначення вагових коефіцієнтів кожного з них у кожної групи.

На основі всіх перерахованих вище показників та подальшого визначення їх вагових коефіцієнтів були складені профілі авторів. Загалом у профіль автора увійшли чотири основні групи згідно з досліджуваними методами. Кожна з груп включає перелік показників з індивідуальними ваговими коефіцієнтами для кожного. Таким чином, для кожного автора було визначено перелік показників, які найточніше відображають його авторський стиль та дозволяють визначати схожі елементи в текстах контрольної вибірки.

Наприклад вектора профілю Т.Г. Шевченко на основі стем, створений на основі словника the Large Electronic Dictionary of Ukrainian (VESUM) [51]:

X′ =[…а, аа, аб, абатів, абатівськ, абатств, абатськ, абет, абетк, аби, аби-аби, абиде, абиколи, абикуди, аби-но, абискільки, абись, аби-то, абич, …].

Всього у векторі 7239 використовуваних у тексті стем. Як видно з отриманих даних, кількість тем для аналізу не менша, що також вимагатиме подальшого скорочення та вибірки найбільш інформативних з них.

Їхні вагові коефіцієнти для профілю Т.Г. Шевченка:Y' =[…0.91, 0.12, 0.55, 0.08, 0.18, 0.82, 0.9, 0.85, 0.99, 0.89, 0.17, 0.86, 0.38, 0.99, 0.42, 0.58, 0.98, 0.62, 0.43, 0.34, …].

І при роботі з правилами при створенні профілю всі правила, отримані в процесі аналізу творів навчальної вибірки були зібрані в єдину базу і для кожного з них було також знайдено вагові коефіцієнти. Загальна кількість правил становила 6946, далі наведено приклад вектора з ваговими коефіцієнтами для них: X' =[…0.35, 0.88, 0.25, 0.44, 0.21, 0.6, 0.41, 1, 0.08, 0.2, 0.72, 0.21, 0.86, 0.49, 0.62, 0.12, 0.54, 0.14, 0.12, 0.24, …].

Кількість правил дещо менша, проте все ще вимагає виділення найбільш важливих та інформативних для коректного визначення авторства з найменшими витратами ресурсів.

Для повторного експерименту профіль кожного автора було зменшено для кожної групи показників. Були відкинуті показники з найменшими ваговими коефіцієнтами для кожної групи з метою скорочення витрат у часі та обчислювальної потужності комп'ютера.

Представлені твори наступних авторів: IB – І. Багряний, AV – А. Вишня, MV – М. Вовчок, AD – А. Довженко, НК – Н. Квітка-Основ’яненко, РМ – П. Мирний, VN – В. Нестайко, VP - В. Підмогильний, IF - І. Франко, МК - М. Хвильовий.

Отримані результати. У роботі з контрольною вибіркою щодо авторства тексту з урахуванням профілю автора було отримано такі результати табл. 3.18. На основі представлених даних при роботі з профілем автора кількість творів з правильно визначеним авторством у контрольній вибірці склала 54 роботи з 60.

Досліджуваний метод дозволив визначити авторство більшості текстів вірно, твори таких авторів як Вишня, Довженка, Мирний, Нестайко та Підмогильний були визначені усі без винятків. У той час як при порівнянні профілю наступних авторів – Багряного, Вовчка, Квітки-Основ'яненка, Франка та Хвильового – один із творів було визначено не вірно і показало велику схожість із профілем іншого автора вибірки.

Таблиця 3.18 – Визначення авторства текстів повного профілю автора

Дійсний	*Встановлений*	*Дійсний*	*Встановлений*	*Дійсний*	*Встановлений*	*Дійсний*	*Встановлений*
IB	IB	MV	MV	PM	PM	VP	VP
IB	IB	MV	MK	PM	PM	VP	VP
IB	IB	MV	MV	PM	PM	VP	VP
IB	IB	AD	AD	PM	PM	IF	IF
IB	IB	AD	AD	PM	PM	IF	IF
IB	IF	AD	AD	PM	PM	IF	IF
AV	AV	AD	AD	VN	VN	IF	IB
AV	AV	AD	AD	VN	VN	IF	IF
AV	AV	AD	AD	VN	VN	IF	IF
AV	AV	HK	HK	VN	VN	MK	MK
AV	AV	HK	HK	VN	VN	MK	MK
AV	AV	HK	MK	VN	VN	MK	MK
MV	MV	HK	HK	VP	VP	MK	MK
MV	MV	HK	HK	VP	VP	MK	MK
MV	MV	HK	MK	VP	VP	MK	IF

При аналізі отриманого результату деяку схожість стилів у двох творах показали Багряний та Франко, а також можна стверджувати, що стиль Хвильового найчастіше перегукується зі стилями інших авторів: у 3 випадках із 6.

За винятком із переліку показників найменш значущих кожному за автора. Таким чином, кількість 4-грамів у профілі знизилася на 1750, стем на 1448 і правил на 1390, що склало 20% у кожному з класів. Під час роботи з оптимізованими векторами було отримано такі результати табл. 3.19.

В результаті проведення експерименту із застосуванням генетичного алгоритму та визначенням найкращого рішення було отримано такі результати: з

60 текстів контрольної вибірки авторство 54 з них було встановлено правильно, що становило загалом 90%.

Для порівняння у попередніх роботах та застосуванні кожного із зазначених методів окремо були отримані наступні результати. Найкращий показник – 91% збігів авторства текстів – було отримано під час роботи з 4-грамами. Робота з основами слів із застосуванням словників та стемінгу дала результат 88%.

Таблиця 3.19 – Визначення авторства текстів скорочений профіль автора

Дійсний	*Встановлений*	*Дійсний*	*Встановлений*	*Дійсний*	*Встановлений*	*Дійсний*	*Встановлений*
IB	IB	MV	MV	PM	PM	VP	VP
IB	IB	MV	MK	PM	PM	VP	VP
IB	IB	MV	MV	PM	PM	VP	VP
IB	IB	AD	AD	PM	PM	IF	IF
IB	IB	AD	AD	PM	PM	IF	IF
IB	IF	AD	AD	PM	PM	IF	IF
AV	AV	AD	AD	VN	VN	IF	IB
AV	AV	AD	AD	VN	VN	IF	IF
AV	AV	AD	AD	VN	VN	IF	IF
AV	AV	HK	HK	VN	VN	MK	MK
AV	AV	HK	HK	VN	VN	MK	MK
AV	AV	HK	MK	VN	VN	MK	MK
MV	MV	HK	HK	VP	VP	MK	MK
MV	MV	HK	HK	VP	VP	MK	MK
MV	MV	HK	MK	VP	PM	MK	IF

Як видно, поєднання різних підходів і методів суттєво не покращило результат, проте, дозволило врахувати додаткові особливості тексту завдяки роботі з граматиками.

Виходячи з отриманих даних найбільш успішними методами роботи з текстом є 4-грами - робота з ними є середньою за витратою ресурсів та часу, щодо інших методів і дає найкращий результат. А також робота зі стохастичними граматиками, завдяки відображенню особливостей побудови автором фраз та речень, проте цей метод потребує значних обчислювальних та часових ресурсів.

Результат роботи зі стемами та словниками показує їх меншу інформативність. Беручи до уваги великі витрати даних методів у обчисленнях та

часі, методи є найбільш витратними та найменш інформативними серед усіх використовуваних. За винятком найменш значущих показників і як наслідок скорочення їх кількості отриманий результат становив 52 твори з правильно встановленим авторством, що хорошим результатом – 87% вірності визначення.

Даних підхід дозволив відчутно знизити складність та час розрахунку, тоді як результат знизився не значно.

Даний підхід дозволив отримати ефективний профіль автора з урахуванням різних особливостей його особистої мови, від використання окремих слів і особливостей побудови речень. Отримані результати демонструють ефективність комплексного підходу, що забезпечує кращі результати порівняно з підходами, що враховують окремі аспекти авторського стилю.

3.5 Експеримент зі встановлення ефективності конструктивно-продукційної моделі побудови структури речень при роботі з технічними текстами

Мета експерименту. Встановлення ефективності конструктивно-продукційної моделі побудови структури речень при роботі з технічними текстами, за методом та моделями, представленими у підрозділі 2.5.

Експериментальна база. 16 текстових файлів у форматі docx, що є документацією до дипломних проектів ОКР «Бакалавр» за напрямом 6.050103 «Програмна інженерія» ДНУЗТ–2018 (розміром 0,7 Мб – 27,3 Мб). Кожен файл містить структурні розділи (28–33 шт.), кожен розділ виділено в окремий txt-файл. Загальна кількість текстів (файлів) розділів – 509. Технічні характеристиками ПК не впливають на результати експерименту.

Методика експерименту. Попередньо визначені конструктивно-продукційні моделі процесів визначення авторства текстів та їх програмні реалізації застосовані для експериментальної перевірки гіпотези щодо можливого статистичного зв’язку між результатами рішення відповідних задач: завдання виявлення запозичень та завдання встановлення авторства тексту відповідно до стилістики та інших особливостей авторського тексту.

Далі представлена перша частина експерименту, а саме послідовність виконання пошуку запозичень у природньомовних текстах за допомогою конструктора графів складається з наступних кроків (рис. 3.9).

Експеримент складається з трьох логічних частин:

– визначення відсотку запозичень у тексті з використанням моделі графового представлення тексту [71, 116];
– визначення відсотку запозичень шляхом аналізу авторського стилю;
– обчислення коефіцієнту кореляції результатів п. 1 та 2.

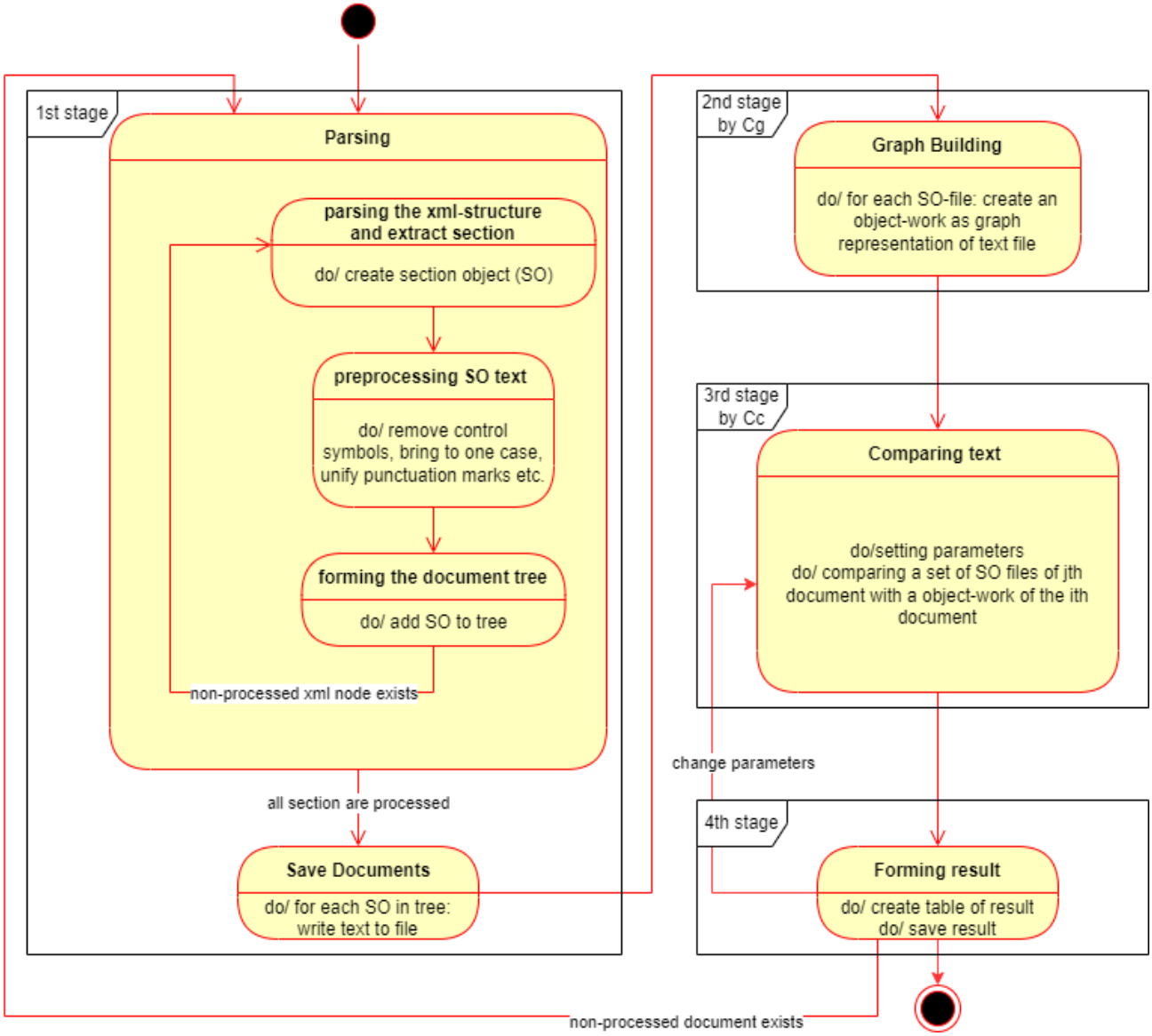

Рисунок 3.9 – Послідовність виконання пошуку запозичень у природньомовних текстах за допомогою конструктора графів

Перший крок експерименту є спільним для обох частин, це автоматизований аналіз структури документу, що виконується на основі розбору xml-структури його файлу, за якою заголовки, що оформлені за допомогою вбудованих стилів заголовків, визначають межі розділів; формування txt-файлів, що містять тексти окремих розділів. При формуванні файлів тексти розділу проходять попередню обробку: видалення керуючих символів, приведення до одного регістру, уніфікація пунктуаційних знаків тощо.

Другим та третім кроком є побудова графового представлення текстів файлів розділів і-го документа та встановлення параметрів та порівняння набору txt-файлів j-го документу з графовим представленням розділів і-го документа відповідно.

Останнім етапом першої частини експерименту є формування зведеної таблиці результатів (табл. 3.20) та завдання нових значень параметрів порівняння для повторення пунктів 3-4 та перехід до (j+1)-го документу.

При роботі з конструктором для відображення побудови структури речення відбувається перетворення тексту з txt-файлу у текований текст з зазначенням частин мови, числа та роду за допомогою першого конструктору-перетворювача C_P; формування правил стохастичного конструктору на основі тегованого тексту другим конструктором-перетворювачем C_T. За допомогою конструктора-вимірювача C_E розрахунок подібності двох стохастичних конструкторів, що відображають синтаксичну структуру текстів, що порівнюються. Останнім кроком є формування зведеної таблиці результатів.

Отримані результати. У табл. 3.20 представлено результати послідовного порівняння відповідних розділів двох дипломів між собою (Р1, Р2,…Р30) з використанням двох описаних вище методів у вигляді відсотку співпадіння між ними.

Через відмінності двох підходів до порівняння – використання речень у першому та послідовностей слів без урахування речень у другому – конструктор графів працював з різними параметрами порівняння, що являє собою тип

фрагменту та його мінімальна довжина, при яких фрагмент може вважатися запозиченим (3-7 відповідно).

Результат роботи конструктора-вимірювача на основі синтаксичної структури речень у двох відповідних розділах розташовано у рядку «речення» табл. 3.20 та приведено до відсоткового вигляду

Таблиця 3.20 – Результати порівняння файлу з використанням конструктора графів та конструктора структури речень

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
речення	**0.00**	**23.00**	**45.00**	3.00	17.00	**62.00**	**0.00**	**42.00**	**6.00**	2.00
3 слова	**0.66**	27.43	62.79	5.56	21.43	65.04	**0.00**	**45.45**	**5.48**	6.67
4 слова	**0.00**	**25.66**	55.81	0.00	0.00	**63.69**	**0.00**	**45.45**	**5.48**	6.67
5 слів	**0.00**	**25.66**	**46.51**	0.00	0.00	**63.18**	**0.00**	**45.45**	**6.85**	0.00
6 слів	**0.00**	**22.71**	**46.51**	0.00	0.00	**62.54**	**0.00**	0.00	0.00	0.00
7 слів	**0.00**	**20.94**	**46.51**	0.00	0.00	**62.54**	**0.00**	0.00	0.00	0.00
	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
речення	**0.00**	60.00	**3.00**	**16.00**	**4.00**	**25.00**	**5.00**	**100.00**	**39.00**	2.00
3 слова	24	35.73	**2.07**	**15.76**	9.69	41.38	4.51	**100.00**	**38.71**	29.63
4 слова	**0.00**	68.43	**3.45**	**15.76**	**5.2**	0.00	4.51	**100.00**	**38.71**	29.63
5 слів	**0.00**	31.86	**3.45**	13.79	**0.00**	**24.14**	4.51	**100.00**	**38.71**	0.00
6 слів	**0.00**	30.47	0.00	13.79	**4.62**	15.52	4.51	**100.00**	**38.71**	0.00
7 слів	**0.00**	28.81	0.00	13.79	**3.93**	15.52	0.00	**100.00**	0.00	0.00
	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
речення	**1.00**	**0.00**	**19.00**	**68.00**	**0.00**	**0.00**	**56.00**	**17.00**	**14.00**	**0.00**
3 слова	**0.42**	2.00	20.95	22.92	**0.00**	**0.00**	60.37	**15.15**	24.24	16.67
4 слова	**0.76**	3.33	**18.1**	19.79	**0.00**	**0.00**	**58.74**	**15.15**	**15.15**	**0.00**
5 слів	0.00	3.33	**18.1**	**73.95**	**0.00**	**0.00**	**55.7**	**15.15**	**15.15**	**0.00**
6 слів	0.00	**0.00**	**18.1**	19.79	**0.00**	**0.00**	52.99	0.00	0.00	**0.00**
7 слів	0.00	**0.00**	**18.1**	19.79	10.67	**0.00**	47.77	0.00	0.00	**0.00**

.Результат порівняння розділів на основі графого представлення розташовані у рядках 3-7 слів у табл. 3.20, в залежності від обраної довжини послідовностей слів для порівняння. Для наочності близький відсоток подібності розділів у таблиці виділено. Отримання нульової подібності деяких розділів зазвичай зумовлено їх надто малим розміром, щоб достовірно відображати подібність.

В даній роботі: тип фрагменту – слово, мінімальна довжина: від 3 до 7 слів. Така довжина зумовлена результатами досліджень [8], дані якого частково наведені в табл. 3.21.

Оскільки експериментальна база даного дослідження містить тексти наукового стилю, які в основному мають повну граматичну основу та другорядні члени речення, то мінімальною довжиною речення обрано три слова.

Щодо максимальної довжини, виходячи з даних табл. 3.21, доцільним брати 7–9. Проте паралельне виконання другої частини експерименту вказують на достатність розгляду максимуму, рівного семи.

Таблиця 3.21 – Довжини речень, які вживаються найчастіше

Довжина в словах	*Драматургія, %*	*Художня проза, %*	*Поезія, %*
1-3	49,73		
4-6	29,07	18,6	18,87
7-9	12,14	18,65	23,02
10-12		18,33	18,33

Отримані результати подібності для 16 дипломних робіт було порівняно і отримано коефіцієнт кореляції для результатів роботи двох описаних підходів. Порівняння було проведено з урахуванням різних довжин послідовності слів, кількістю від 3 до 7 та отримано наступні результати.

Таблиця 3.22 – Кореляційна подібність результатів отриманих з використанням конструктора структури речень та конструктора графів з 5 словами поспіль

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		96	91	84	77	84	89	89	97	81	81	83	94	85	97	82
2	98		95	88	86	96	80	90	86	86	100	75	77	85	79	83
3	97	99		82	81	94	98	79	88	86	76	83	91	79	96	83
4	90	98	80		89	75	86	77	76	80	97	84	98	94	87	99
5	88	89	95	95		82	81	89	99	94	91	86	85	75	78	87
6	95	89	85	82	86		99	91	98	83	97	89	96	77	98	88
7	98	84	91	84	99	82		89	87	87	99	78	86	91	81	79
8	88	93	99	75	84	83	91		81	93	89	76	89	89	80	98
9	90	89	92	86	87	91	99	86		84	87	97	75	96	88	87
10	97	99	87	86	85	75	87	98	100		98	93	75	98	79	83
11	91	93	79	87	92	90	90	76	86	89		95	84	77	93	82
12	90	88	91	89	83	91	98	86	76	93	75		91	89	94	84

13	88	99	78	80	85	93	95	83	91	96	95	98		94	84	78
14	94	99	87	91	95	80	90	79	98	81	86	82	97		98	75
15	96	87	94	86	94	78	98	79	99	87	96	88	85	82		98
16	99	95	76	93	79	89	90	76	90	82	96	76	91	75	78	

Для 3 слів поспіль середнім значенням коефіцієнту кореляції став 0.00053, що є незадовільним результатом та демонструє велику розбіжність результатів поміж двох застосованих методів.

При використанні послідовності слів довжиною 4 достовірність результатів значно покращилась – середнє значення становить 0.82, що дозволяє сказати що аналіз починаючи з 4 слів є достовірним та відображає реальний стан речей.

При подальших експериментах з використанням довжини слів 5 та 6 отримані результати також відображають доцільність використання саме цих довжин послідовностей слів як найбільш інформативних. Середнє значення коефіцієнтів кореляції становить 0.88 для розрахунків з довжиною послідовності 5 та 0.82 для довжини 6 слів.

Результати роботи з послідовністю слів довжиною 7. Результат подібний до роботи з 3 словами поспіль і вказую на недоцільність їх використання. Середнє значення кореляційного коефіцієнта становить 0.000531, що є незадовільним результатом і говорить про сильну розбіжність двох методів.

Загальним результатом можна вважати достатню кореляційну подібність двох методів та виявлення потрібних довжин послідовностей для достовірного відображення авторського стилю.

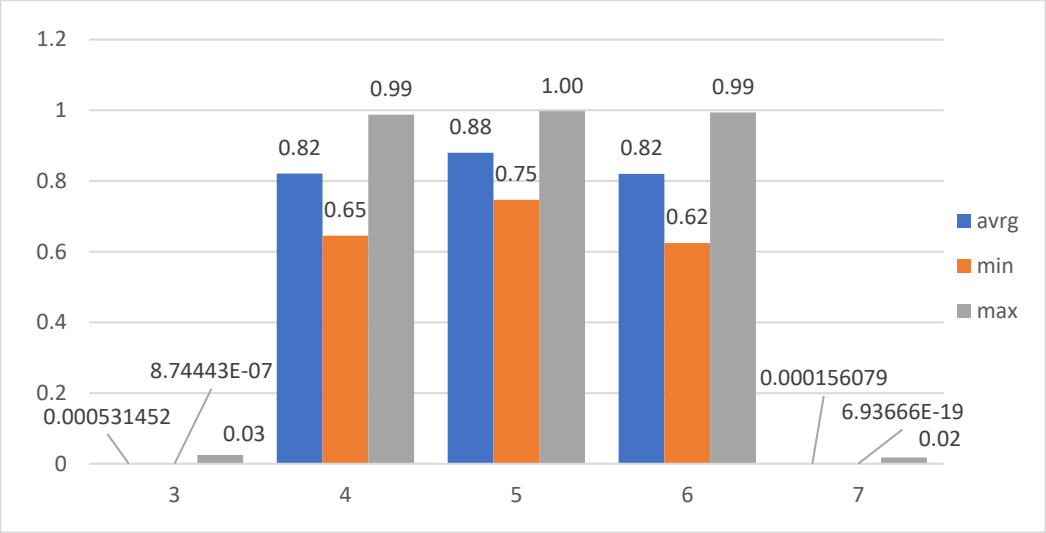

Рисунок 3.10 – Максимальні та мінімальні значення кореляційної подібності двох методів для 16 дипломних робіт з використання послідовностей довжиною 3-7

Дослідження виконано на технічних текстах з програмування. Вважаємо, що метод може застосовуватись і для текстів з інших технічних галузей, але це положення слід підкріпляти відповідними експериментами.

При малих об’ємах тексті спостерігалась слабка кореляція результатів з виявлення запозичень та встановлення авторства текстів, що потребує подальших досліджень. Чітких меж між малими та великими обсягами текстів не існує. За результатами експериментів встановлено, що для задовільного результату мінімальний обсяг текстів має бути 10 символів та складатися з 5 речень.

При виборі взірців тексту автору слід враховувати що стиль автора може змінюватись через плив часу, різну тематику текстів та зміну загальновживаних шаблонів формування текстів.

Вважаємо що запропонований метод встановлення авторства текстів може бути поширений та ефективно використовуватись для різних слов’янських мов. Інші мови, наприклад англійської, де менша атрибутика слів, яка компенсується шаблонами синтаксису речень, а також більш формальними вимогами до нього, значно послаблюють можливості запропонованих методів.

Представлений метод встановлення авторства текстів може використовуватись також для виявлення підозрілих щодо наявності великого

обсягу запозичень. Що може слугувати приводом для подальшої, більш ретельної перевірки програмними засобами або з залученням експертів.

Для перевірки тексту на наявність плагіату необхідно мати великий банк текстів інших авторів для виявлення цих запозичень, що може бути складним через постійне збільшення кількості матеріалів у вільному доступі та різноманітність форм та форматів їх представлення. На відміну від відомих програм з виявлення запозичень та встановлення авторства, запропонований підхід обмежуються наявністю відносно невеликого обсягу текстів автора та не потребує наявності великого банку текстів.

У ході роботи з пояснювальними записками до дипломів студентів-програмістів було встановлено, що підхід дозволяє працювати лише з природньомовним текстом. Метод не спрацював при роботі з розділами, що включають у себе код програм, який не було можливо опрацювати.

У ході експерименту підтверджено гіпотезу щодо високого зв’язку між задачами, методами їх рішення та результатами щодо встановлення авторства технічних текстів та виявлення запозичень. Встановлено, що кореляційне відношення між результатами може становити більше 90%.

Розроблена конструктивно-продукційна модель встановлення авторства тексту на основі аналізу особливостей та закономірностей авторського стилю формування речень. Сутність моделі полягає у формалізації представлення синтаксису речень автора множиною правил підстановки з ймовірнісним навантаженням. Отримані результати свідчать що запропонований метод має високу ефективність у порівнянні з методами, що раніше використовувалися.

Розроблено модель технічних текстів з врахуванням авторського стилю, завдяки відображенню унікальних стилістичних та мовних особливостей власної мови автора дозволило значно спростити процес виявлення запозичень та встановлення авторству через необхідність використання лише однієї роботи автору замість цілого корпусу текстів – можливих джерел запозичень.

Результати експериментів визначили значення раціонального параметру – мінімальну довжину текстових фрагментів (у кількості слів) яку слід вважати

запозиченням, він дорівнює п’яти словам. Це слід вважати рекомендацією для використання будь-яких програм виявлення запозичень.

Запропонований метод може використовуватись як для вирішення проблем пошуку запозичень, так і для встановлення вірогідного авторства тексту.

Висновки по третьому розділу

У розділі викладені результати дослідження ефективності застосування адаптованого та розробленого підходу та інструментарію для визначення авторства природньомовних текстів.

Усі результати досліджені показали що, методи дали позитивний результат і можуть використовуватися для визначення авторства текстів. Найкращий показник – 91% збігів авторства текстів – було отримано під час роботи з 4-грамами та 90% при роботі з конструктивно-продукційною моделлю мови автора.

Запропоновані методи можуть використовуватись як для вирішення проблем пошуку запозичень, так і для встановлення вірогідного авторства тексту.

За матеріалами розділу опубліковано роботи [18, 19, 20, 21, 22, 23, 107, 108, 109, 110, 111, 112, 113].

РОЗДІЛ 4
РОЗРОБЛЕНИЙ ПРОГРАМНИЙ ІНСТРУМЕНТАРВІЙ З РЕАЛІЗАЦІЇ ЗАСТОСОВАНИХ МЕТОДІВ ТА МОДЕЛЕЙ

4.1 Визначення кола основних задач програмного інструментарію

Було визначено коло задач, що повинні вирішуватись програмним забезпеченням задля отримання бажаного результату.

Основна функція передбачає отримання списку конкретних індикаторів, що будуть слугувати чисельним відображенням індивідуальних особливостей мовлення автора. Ці показники розраховуються відповідно до попередньо визначеної методології, яка є невід'ємною частиною дизайну програми. Ця функція є ключовою, оскільки забезпечує розрахунок та компіляцію широкого спектру точних даних або показників на основі заданих параметрів.

Ще однією важливою функцією програми є створення вектору-образу тексту, що досліджується. Цей процес передбачає отримання ряду різних показників шляхом проведення зазначених у попередніх розділах розрахунків, їх нормалізацію та подальше представлення у векторній формі. Створення цього вектору-образу тексту, що досліджується, є ключовим кроком, оскільки це дозволяє відобразити індивідуальні особливості та закономірності у стилі мовлення автора наданого тексту, перетворюючи його у формат, придатний для порівняння та аналізу.

Наступним кроком програма використовує вектори-образи текстів для порівняння один з одним чи зі заздалегідь визначеним стандартом – вектором-образом середнього значення кожного з показників для певного автора, що отримується на основі навчальної вибірки з його робіт. Шляхом аналізу програма виявляє розбіжності або подібності між згенерованим вектором і вектором еталоном, щоб на його основі зробити висновок щодо близькості стилю написання тексту, що досліджується, художньому стилю одного з авторів.

Тож, програма має попередньо оброблювати обраних текст для можливості проведення подальших розрахунків, отримувати певний набір показників з урахуванням обраних параметрів. До переліку показників увійшли 3 їх класи показників: складності сприйняття тексту, рекурентні та частотні характеристики входження певних літер або їх послідовностей. Та окремо побудова структурних правил для опису особливостей написання речень автором з розрахунком частоти появи певної частини мови на певному місці.

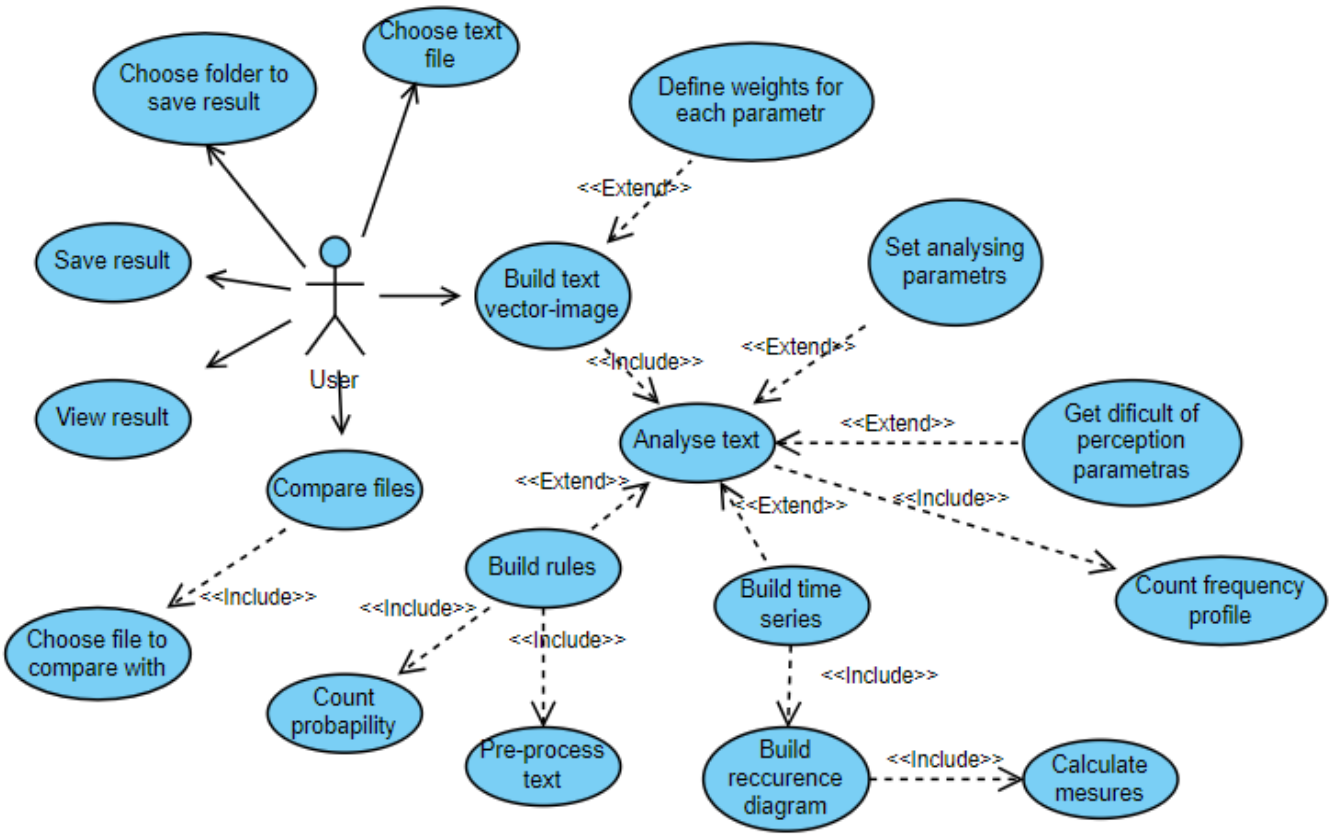

Рисунок 4.1 – Діаграма варіантів використання

У увесь спектр задач, що вирішує програма відображений на діаграмі варіантів використання рис. 4.1.

4.2 Логічне розбиття програми на частини

Через велику кількість задач та підзадач, а також загальну складність розрахунків у числовому та обчислювальному еквіваленті було вирішено розбити

задачі на групи. Для надання користувачеві можливості вибору методу аналізу тексту та відбору даних для включення у вектор-образ, налаштування вагових коефіцієнтів та безпосередньо порівняння, загальний інструмент було поділено на 3 логічні блоки згідно їх функціоналу.

Загальний вигляд та логіка зв'язку програм, логіку розподілення завдані між ними та послідовність використання розроблених інструментів продемонстровано на рис. 4.2.

Діаграма візуально ілюструє комплекс різноманітних інструментів, кожен з яких розроблений для виконання певних ролей у системі.

4.2.1 Функціональність реалізована пакетом Attribution

Серед цих інструментів виділяється пакет атрибуції (Attribution), що включає три різні програми, розроблені для безпосередньої взаємодії з текстовими даними. Ці програми використовують різні методології для обробки вхідного тексту, кожна з яких зосереджується на різних підходах до аналізу тексту та розрахунку атрибутів.

Reccurence diagram – частина відповідальна за розрахунки частоти входження певної текстової одиниці (літери, певної їх послідовності, стеми, тощо), побудови частотної діаграми та частотного ряду на основі частотного профілю. Наступним кроком будується рекурентна діаграма та розраховуються рекурентні показники. Результат зберігається у вигляді файлу з розрахованими даними та зображеннями самого частотного ряду та рекурентної діаграми на його основі.

Text perception difficulty – частина що розраховує специфічні показники щодо складності сприйняття тексту та загальну статистику як математичне очікування та середнє квадратичне відхилення для заданих показників.

Building rules – частина, що реалізує перетворення тексту на набір правил, що відображають структуру побудови автором речень у обраному тексті. Відповідальна за перетворення тексту з txt-файлу у тегований текст з зазначенням частин мови, числа та роду та є реалізацією першого конструктору-

перетворювача C_P та другого конструктора-перетворювача C_T, що безпосередньо формує правила стохастичного конструктору на основі тегованого тексту.

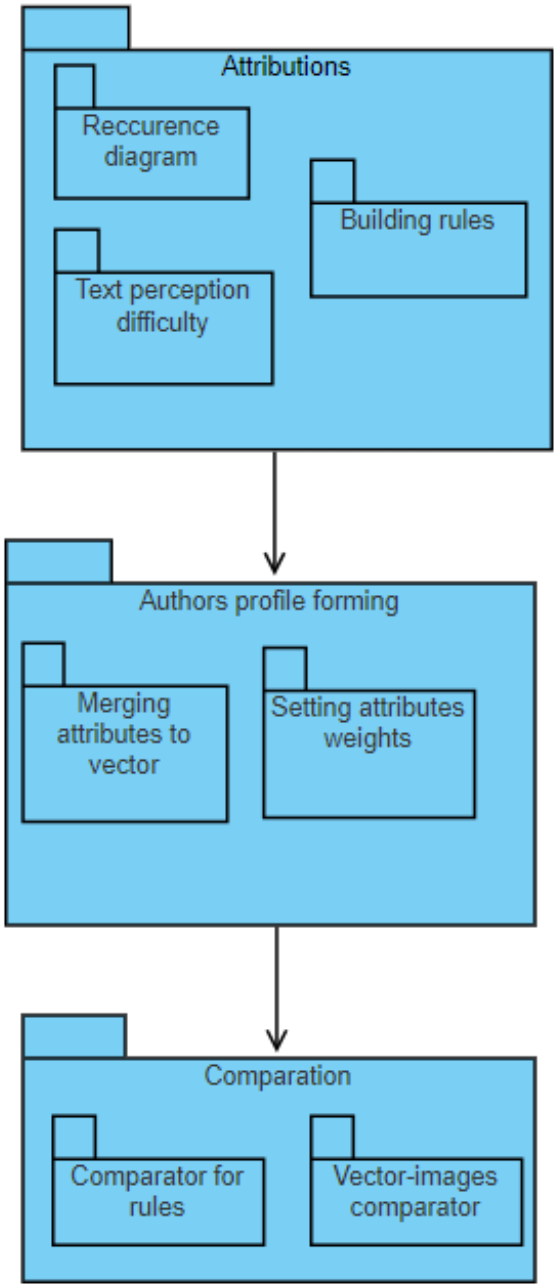

Рисунок 4.2 – Загальна схема розробленого інструментарію

Кожен компонент у пакеті атрибуції відіграє ключову роль у комплексному аналізі текстового вмісту та особливостей його побудови. Ці компоненти виконують відмінні функції, наголошуючи на окремих обчисленнях і отриманні атрибутів, призначених для відображення конкретних характеристик, притаманних тексту. За допомогою алгоритмів та методології розробки ці

програми можуть виконувати обчислення окремо одна від одної та отримувати як повний перелік атрибутів із введеного тексту, так і окремі данні за запитом.

Усі дані, отримані у результаті виконання обчислень організовуються та зберігаються в структурованому форматі у окремому файлі. Це виконує подвійну мету: по-перше, як наочне представлення для обчислених атрибутів, а по-друге, як ресурс для майбутнього використання в різних операціях, таких як створення профілю автора або порівняння окремих показників.

4.2.2 Функціональність реалізована пакетом Authors profile forming

Пакет Authors profile forming вирішує задачу безпосереднього створення профілю автора та складається з двох частин: Merging attributes to vector Setting та attributes weights.

Пакет формування профілю автора виконує роботу над загальним профілем автора – об’єднує різні атрибути у певному порядку, нормалізує їх та підбирає вагові коефіцієнти відповідно до інформативності певного атрибута для загального профілю. Для формування профілю автора усі отриманні данні збираються в єдиний вектор, якому у відповідність формується вектор вагових коефіцієнтів, що змінюються в процесі їх підбору. Цей пакет також відповідальний за зменшення профілю автора – вилучення з профілю найменш інформативних показників згідно ваговим коефіцієнтам задля скорочення необхідних на проведення розрахунків часу та обчислювальних ресурсів.

Нижче представлено загальну схему логіки створення профілю автора, розрахунку вагових коефіцієнтів параметрів та зменшення вектору для використання меншої кількості атрибутів при розрахунках (рис. 4.3).

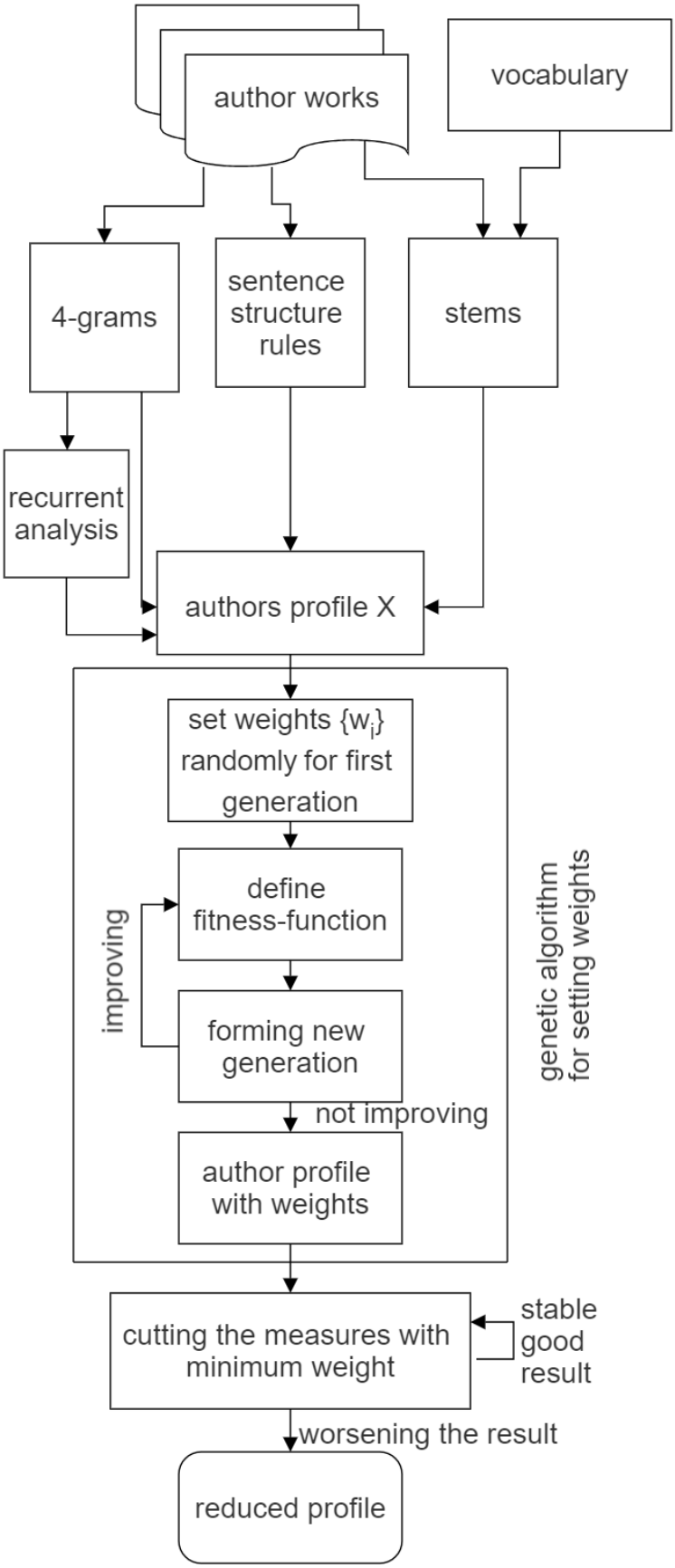

Рисунок 4.3 – Загальна схема формування профілю автора

Перша частина створює профіль автора відповідно обраним класам показників, нормує їх задля коректності проведення подальшого порівняння векторів та також зберігає у файл для можливості подальшого використання.

Друга частина проводить розрахунки за допомогою реалізації генетичного алгоритму для підбору вагових коефіцієнтів кожного з показників для покращення подальшого результату порівняння та можливості визначення набору найбільш значущих серед них для стилю мовлення певного автора.

Також саме у другій частині виконується скорочення профілю автора згідно найменшим ваговим коефіцієнтам.

Пакет розбито на дві окремі частини задля можливості їх як одночасного, так і окремого використання.

4.2.3 Функціональність реалізована пакетом Comparation

У пакет Comparation відповідальний за порівняння окремих текстів один з одним, чи з загальним профілем автора.

Етап порівняння та ідентифікації авторства досліджуваного тексту виступає як окремий і ключовий етап у системі. На цьому етапі програма працює безпосередньо з багатогранними елементами, такими як вектори профілю автора, вектори обчисленої вагових коефіцієнтів та вектор-образ, що представляє текст, що досліджується.

Пакет розбито на дві частини згідно з задачами порівняння набору правил чи безпосередньо певних показників (Comparator for rules та Vector-images comparator відповідно). Це зумовлено різницею підходів до порівняння, оскільки при роботі з числовими показниками певного класу розраховується відстань до еталону (середньому вектору-образу певного автора чи окремого тексту) для отримання числа, що відображає близькість векторів-образів, тобто особливостей авторського стилю відповідно.

Однак при роботі з переліком правил, що описують особливості побудови речень у тексті, правило розглядається як єдине ціле і проводиться пошук найдовшого ланцюжка співпадінь правил в обох наборах правил (для двох текстів,

чи тексту та збірного набору правил, що характерні для стулю певного автора). У цьому випадку розрахунок проводиться ступеню їх статистичної структурної подоби визначатиметься як добуток мінімальної різниці ймовірностей застосування відповідного правила та сума ступенів подоби усіх співпадаючих правил у двох наборах. У цьому випадку 1 позначатиме повне співпадіння, а 0 – відсутність співпадінь.

Друга частина пакету, що відповідальна за обчислення подібності структури речень у тексті реалізує розроблений у Розділ 2 конструктора-вимірювача C_E.

4.3 Послідовність виконання пошуку співпадінь у природньомовних текстах на прикладі конструктора структури речень

Наведемо приклад, що демонструватиме процес розрахунку певного класу показників та порівняння текстів на прикладі роботи з особливостями подови автором речень у тексті.

Першим етапом є попередній аналіз та обробка тексту для формування правил побудови його речень. Програма веде підрахунок речень у тексті, аналізує частини мови, які входять до нього та формує проміжний результат, що надалі використовується для розрахунку безпосередньо вірогідності утворення речення.

Отримавши набір правил – образ тексту, що відображає усі особливості побудови речень у певному тексті, та зберігається для подальшого використання при порівнянні. Кожному етапу відповідає один зі створених конструкторів. Загальна схема процесу представлена на рис. 4.4.

Послідовність виконання пошуку співпадінь у природньомовних текстах складається з наступних етапів:

1) автоматизований аналіз структури документу, що виконується на основі розбору xml-структури його файлу, за якою заголовки, що оформлені за допомогою вбудованих стилів заголовків, визначають межі розділів; формування txt-файлів, що містять тексти окремих розділів. При формуванні файлів тексти

розділу проходять попередню обробку: видалення керуючих символів, приведення до одного регістру, уніфікація пунктуаційних знаків тощо.

2) перетворення тексту з txt-файлу у тегований текст з зазначенням частин мови, числа та роду за допомогою першого конструктору-перетворювача C_P;

3) формування правил стохастичного конструктору на основі тегованого тексту другим конструктором-перетворювачем C_T;

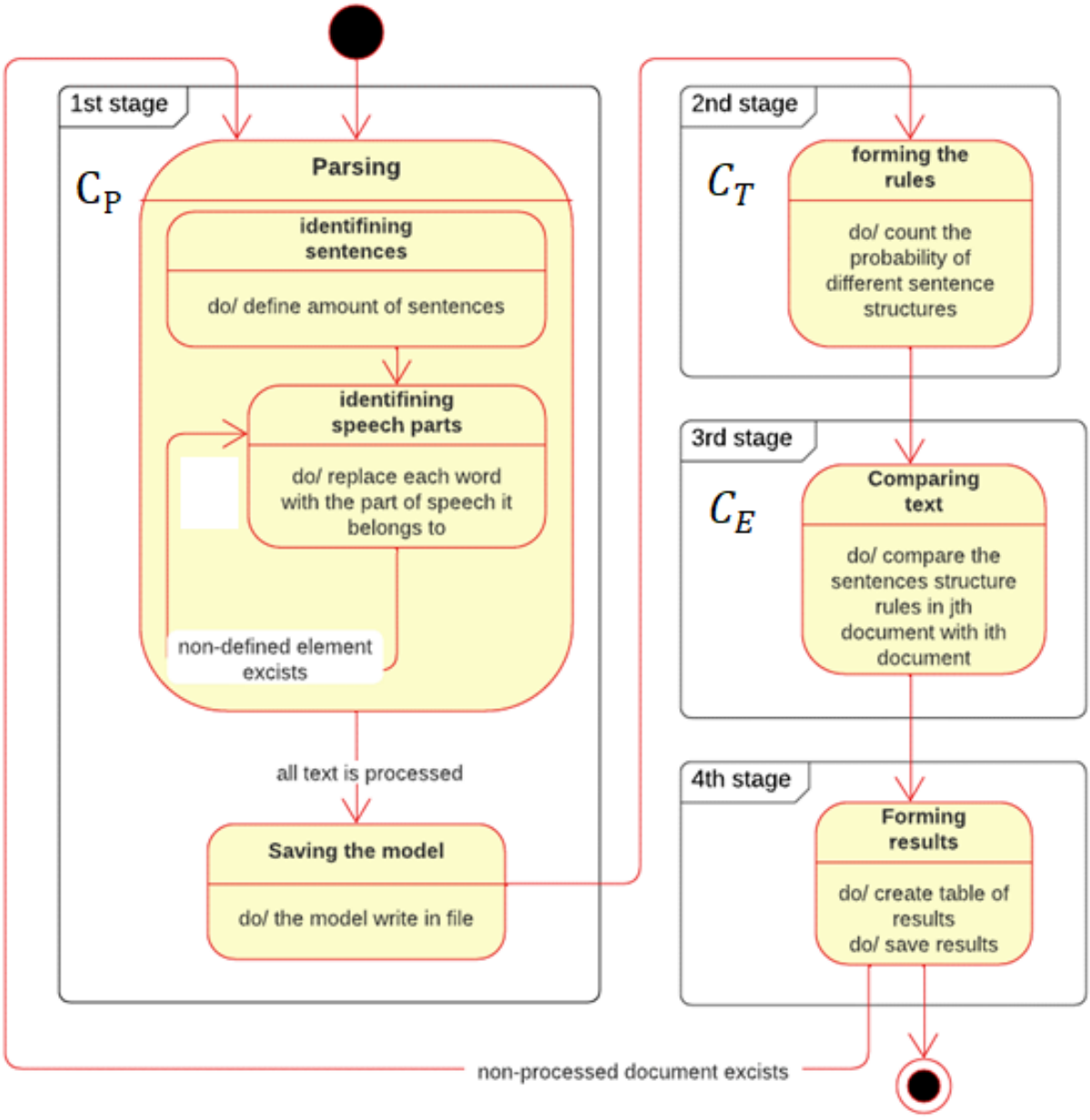

Рисунок 4.4 – Послідовність виконання пошуку співпадінь у природньомовних текстах на прикладі конструктора речень

4) за допомогою конструктора-вимірювача C_E розрахунок подібності двох стохастичних конструкторів, що відображають синтаксичну структуру текстів, що порівнюються;

5) отримання результатів;

У результати виконання наведеної послідовності дій буде отримано число, що відображатиме ступінь близькості двох текстів за особливостями побудови речень.

4.4 Реалізація пакету Attribution

Для реалізації описаних вище методів розроблено відповідний інструментарій. Була створена програма, що аналізує текст, розраховує відповідні параметри та виконує порівняння наборів даних для визначення їх подібності та встановлення найближчого за стилем автора.

Об’єктно-орієнтовна програма має три рівня: рівень представлення, що працює безпосередньо з користувачем та отримує від нього певні дані; рівень логіки, що містить класи відповідні за обробку текстів та виконання відповідних розрахунків та рівень роботи з даними, що безпосередньо працює зі зчитуванням тексту та збереженням отриманих результатів у відповідній формі.

Згідно рис. 4.5 видно, що за розрахунок кожного з різних атрибутів, перелік яких було надано у попередньому розділі у експериментах 1-3, відповідає певний клас. Для формування конструктивно-продукційної моделі текст, попередньо опрацьований аналізується з точки зору частини мови. Надалі формуються речення з розрахунком появи певної частини мови на певному його місці, що наступним кроком перетворюється на набір правил їх побудови з імовірнісними мірами. Таким чином отримується загальна конструктивно-продукційна модель тексту.

Модель включає 14 класів, розглянемо наведену структуру кожного з рівнів більш детально більш детально.

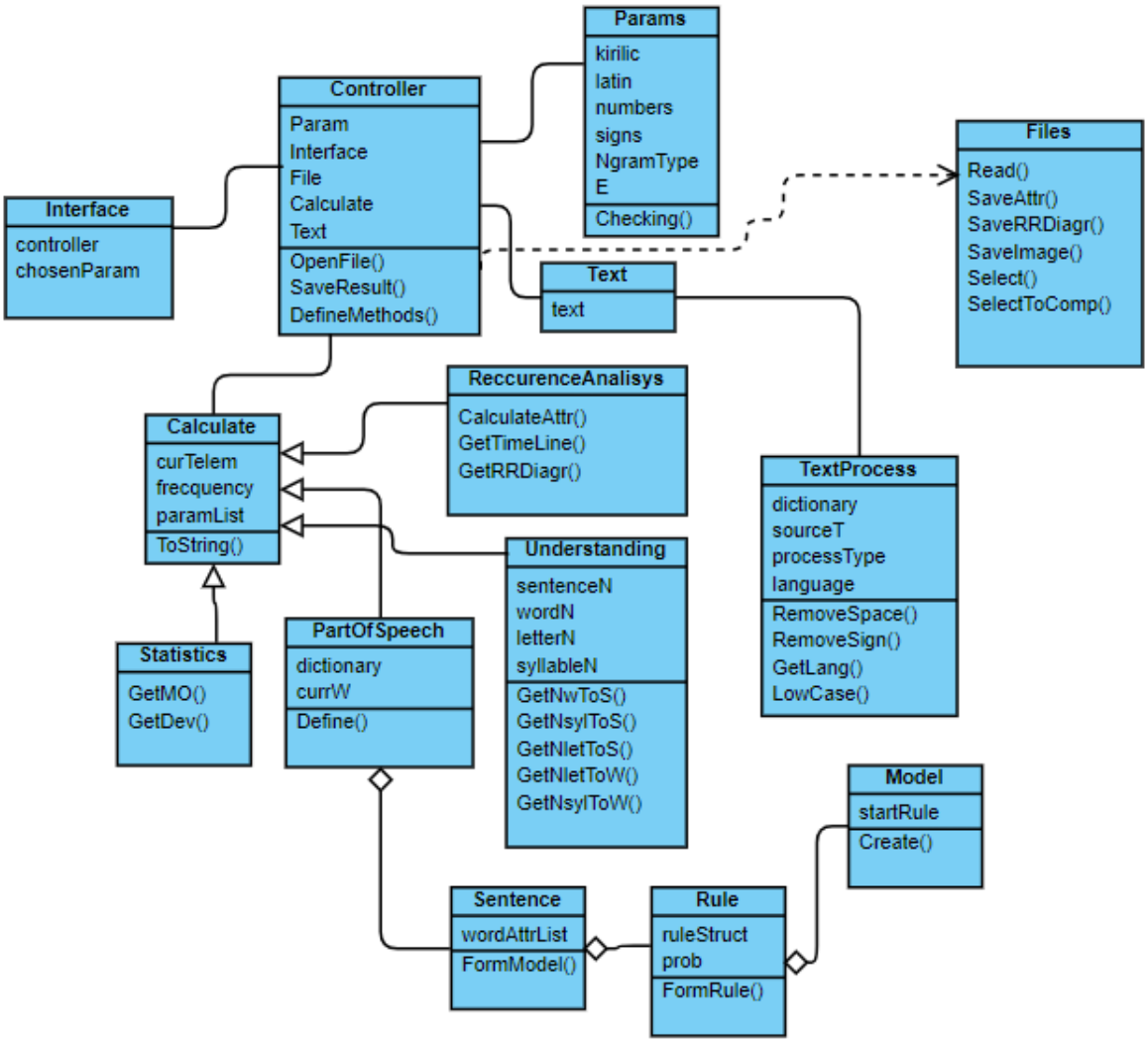

Рисунок 4.5– Об'єктно-орієнтовна модель пакету для розрахунку атрибутів

Рівень представлення: Interface – реалізує інтерфейс користувача та його візуалізацію.

Рівень логіки: Controller – розподілення задач між рівнями, Params – зберігає параметри для розрахунку (клас показників та значення змінних для обчислення), Text – моделює текст, що досліджується, Calculate – збереження результатів обчислень, Statistics, ReccuranceAnalisys, Understanding – розрахунок певного класу показників, TextProcess – попередня обробка тексту для проведення розрахунків, PartOfSpeech – обробка тексту з точки зору частин мови для подальшої побудови правил, Sentence, Rule, Model – реалізують представлення тексту у вигляді послідовностей правил з вірогідністю його спрацювання.

Рівень даних: Files – зчитування тексту з файлу та збереження отриманих результатів.

Вхідними даними для програми є тексти українською мовою, у форматі .txt для їх подальшого аналізування та проведення розрахунків. Вихідними даними – файли, що зберігають окремо зображення діаграм та графіків (частотна діаграма, фазовий простір, рекурентна діаграма, файл з отриманими чисельними даними, файл з правилами).

Висновки по четвертому розділу

В ході роботи над дисертаційним дослідженням розроблено конструктивно-продуктивну модель графового представлення тексту та логіку програм в рамках їх алгоритмічної складової, що дозволило отримати комп’ютерні реалізації моделей. Отримані реалізації дозволяють:

– з метою встановлення авторству природньомовних українських текстів, виконувати порівняння текстових фрагментів;

– формувати вектор-образ тексту для подальшого його використання для встановлення авторства тексту або їх порівняння.

За матеріалами розділу опубліковано роботи [107, 108, 109, 110, 111].

ВИСНОВКИ

У роботі розробляються методи та програмний інструментарій для встановлення авторства природньомовних текстів й виявлення запозичень.

У ході роботи було отримано наступні результати:

1. розроблено модель представлення природньомовниго тексту у вигляді правил стохастичної граматики, метод визначення авторства художніх та технічних текстів на основі цієї моделі та досліджена її ефективність за допомогою репрезентативної вибірки. Модель дозволяє враховувати синтаксичні та стилістичні особливості тексту автора;
2. розроблено конструктивно-продукційну модель для автоматичного будування та порівняння цих правил для встановлення подібності текстів. Данна модель дозволяє оцінювати ступінь подібності текстів на основі роботи з ймовірностями у множинах співпадаючих правил;
3. створено метод визначення авторству на основі декількох груп показників на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою розпізнавання образів. Дозволяє працювати з широким переліком неоднорідних показників, що поєднуються для всебічного охоплення особливостей авторського стилю;
4. запропоновано метод встановлення профілю автора для подальшого його використання у визначенні авторства текстів та пошуку їх подібностей на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою багатокритеріального аналізу та методу направленого випадкового пошуку глобальних екстремумів (генетичного алгоритму) для зменшення числа показників на перелік найбільш значущих для кожного з авторів. З’являється можливість враховувати особливості письма притаманні певному автору та використовувати лише найбільш інформативні показники для кожного з авторів. Подібний підхід дозволить зберігати

лише побудований профіль автора без необхідності витрачати ресурси на роботу з повноцінним корпусом текстів;

5. досліджено ефективності кожної з представлених моделей для знаходження оптимального варіанту вирішення задачі дисертаційної роботи, що дозволило виявити перспективні напрями та підходи до вирішення проблеми встановлення авторству природньомовних текстів;
6. встановлено статистично значимий зв’язок результатів рішення задач виявлення запозичень та встановлення авторству текстів, що дозволить розробити уніфікований інструмент вирішення подібних задач;
7. експериментально встановлено, що застосування стохастичної граматики дозволяє встановити дійсне авторство текстів у 75-80% випадків. З подальшим покращенням результату до 83% за допомогою використання довірчого інтервалу;
8. за допомогою проведення експериментів досліджено ефективність методу визначення авторству на основі декількох груп показників на репрезентативній вибірці як використання загальних даних, так і по відповідним групам. Найкращим став загальний результат, що сягнув 91% правильного встановлення авторства на репрезентативній вибірці;
9. підтверджено гіпотезу щодо високого зв’язку між задачами, методами їх рішення та результатами щодо встановлення авторства технічних текстів та виявлення запозичень. Встановлено, що кореляційне відношення між результатами може становити більше 90%;
10. експериментально підтверджено, що представлена модель з використанням профілю автора для знаходження оптимального варіанту вирішення проблеми встановлення авторства природньомовних текстів дозволив отримати 90% вірного визначення авторства на репрезентативній вибірці.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Алексенко, Світлана Федорівна. Основи лінгвометодології. Суми: СумДПУ, 2019, 33 с.

2. Баєвський, В.С. Лінгвістичні, математичні, симеотичні та комп'ютерні моделі в історії та теорії літератури. Studia philologica, 2001.

3. Бук, С. “Слов’янський досвід укладання частотних словників мови письменника”. Проблеми слов’янознавства, вип. 60, 2011, с. 217-224.

4. Войтенко, К. І. “Функціональний стиль художнього мовлення”. Наукові записки Національного університету Острозька академія. т 1, № 26, 2012, с. 53-56.

5. Глибовець, Андрій, та Точицький Володимир. "Алгоритм токенізації та стемінгу для текстів українською мовою." *Наукові записки НаУКМА. Комп'ютерні науки.* 2017, с 4-8.

6. Духновська, К.К., Страшок, Я.А., Шило, П.В. “Інформаційна технологія для проведення лематизації і стемінгу в україномовних текстах.” *Прикладні системи та технології в інформаційному суспільстві.* (2022): 119-127.

7. Калініна, Інна Володимирівна, та Лісовиченко, Олег Іванович. "Використання генетичних алгоритмів в задачах оптимізації." *Адаптивні системи автоматичного управління* 1.26 (2015): 48-61.

8. Кульчицький, І. М. "Дослідження довжини речення та слова у творах Романа Іваничука." *Вісник Національного університету Львівська політехніка. Серія: Інформаційні системи та мережі* 872 (2017): 139-148.

9. Перебийніс, В.С. *Статистичні методи для лінгвістів: Навчальний посібник.* Вінниця, 2002.

10. Плющ, Марія Яківна. "Граматика української мови. Морфеміка. Словотвір. Морфологія." (2010).

11. Рогушина, Ю. В. “Использование критериев оценки удобочитаемости текста для поиска информации, соответствующей реальным потребностям пользователя.” *Проблеми програмювання*. (2007).

12. Рябишев О.В., Єрохін А.Л., Бахмет А.Г. "Аналіз тональності тексту українською мовою. " *Біоніка інтелекту*. 1.96 (2021): 15–21. doi: 10.30837/ bi.2021.1(96).03.

13. Сушко, С. О., Фомичова Л. Я., та Барсуков Є. С.. "Частоты повторяемости букв и биграмм в открытых текстах на украинском языке." *Захист інформації* 12.3 (2010): 91-98.

14. Хомицька, І. Ю., Теслюк, В. М., Базилевич, І. Б. "Ефективність статистичних критеріїв для визначення стильових характеристик текстів. " *Науковий вісник НЛТУ України*. 33.4 (2023): 90-94. doi: 10.36930/40330413.

15. Хомицька І. Ю., Теслюк В. М., Береговський В. В. "Метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови. " *Науковий вісник НЛТУ України* 29.6 (2019): 140-143. doi: 10.15421/40290627.

16. Хомицька І. Ю., Теслюк В. М.. Береговський В. В. "Математичні метод і модель диференціації фоностатистичних структур авторського стилю. " *Науковий вісник НЛТУ України*. 29.7 (2019): 156-159. doi: 10.15421/40290731.

17. Хомицька І. Ю., Теслюк В. М., Базилевич І. Б., Береговський В. В. "Статистичні моделі та програмні засоби розмежування авторських стилів англійської прози. " *Науковий вісник НЛТУ України*. 30.5 (2020): 135-139. doi: 10.36930/40300522.

18. Шинкаренко, В.І, та Демидович, І.М. “Статистичний та рекурентний аналіз природно-мовних. текстів”. *Збірка тез XII Міжнародної науково-практичної конференції “Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті”*, Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2018, с. 120.

19. Шинкаренко, В.И., та Демидович, И.Н. “Рекуррентный анализ естественно-языковых текстов”. *Збірка тез Всеукраїнської науково-методичної конференції «Проблеми математичного моделювання»*. Дніпропетровський державний технічний ун-т, 2018, с. 40-43.

20. Шинкаренко, В.І, та Демидович І.М. “Використання генетичного алгоритму для покращення визначення авторства природньомовних текстів”. *Збірка тез XIV*

Міжнародної науково-практичної конференції "Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті", Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2020, с. 127.

21. Шинкаренко, В.І, та Демидович І.М. "Показатель структурного сходства естественно языкового литературного текста". *Збірка тез XV Міжнародної науково-практичної конференції "Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті",* Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2021. с. 68.

22. Шинкаренко, В.І, та Демидович І.М. "Застосування формальних стохастичних граматик при визначенні авторству текстів" *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2022, с. 293

23. Шинкаренко, В.І, та Демидович І.М. «Застосування конструктивного моделювання при визначенні авторства текстів» *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2023, с.394.

24. Шинкаренко, Виктор, та Александр Жеваго. "Составление расписания занятий университета на основе конструктивного моделирования." *Radio Electronics, Computer Science, Control* 3 (2019).

25. Addin, O., et al. "A Naïve-Bayes classifier for damage detection in engineering materials." *Materials & design* 28.8 (2007): 2379-2386.

26. Aggarwal, Charu C., та Charu C. Aggarwal. *Machine learning for text: An introduction*. Springer International Publishing, 2018.

27. Alekseev, P.M. "Frequency dictionaries". *Quantitative Linguistik: ein internationales Handbuch. Quantitative linguistics: an international*. 2005. pp. 312-324.

28. Alsaleem, Saleh. "Automated Arabic Text Categorization Using SVM and NB." *Int. Arab. J. e Technol.* 2.2 (2011): 124-128. 2011. 2(2).

29. Altmann, Georg, et al. "Úvod do analýzy textov." *Bratislava, Veda–vydavatelstvo Slovenskej akademie vied* (2003).

30. Barros, Rodrigo Coelho, et al. "A survey of evolutionary algorithms for decision-tree induction." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.3 (2011): 291-312, DOI: https://doi.org/10.1109/TSMCC.2011.2157494.

31. Bensefia, Ameur, et al. "Writer identification by writer's invariants." *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. IEEE, 2002, doi: 10.1109/IWFHR.2002.1030922.

32. Bisikalo, O. V. "Formal methods imagery analysis and synthesis of natural language constructions: monograph." *Vinnitsa: VNTU* (2013).

33. Booth, T. L. "Probability representation of formal languages." *IEEE Tenth Annual Symposium on Switching and Automata Theory*. 1969.

34. Brownlee, Jason. *Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end*. Machine Learning Mastery, 2016.

35. Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175, 1994, c. 161–175.

36. Cheng, Zhou, et al. "TreeNet: Learning Sentence Representations with Unconstrained Tree Structure." *IJCAI*. 2018. 4005-4011.

37. Chomsky, Noam. "On certain formal properties of grammars." Information and control 2.2 (1959): 137-167.

38. Chomsky, Noam. "Three models for the description of language." IRE Transactions on information theory 2.3 (1956): 113-124.

39. Conneau, Alexis, et al. "Supervised learning of universal sentence representations from natural language inference data." *arXiv preprint arXiv:1705.02364* (2017).

40. Dai, Zhuyun, and Jamie Callan. "Deeper text understanding for IR with contextual neural language modeling." *Proceedings of the 42nd international ACM SIGIR*

conference on research and development in information retrieval. 2019. doi:10.1145/3331184.3331303.

41. Damanik, Irfan Sudahri, et al. "Decision tree optimization in C4. 5 algorithm using genetic algorithm." *Journal of Physics: Conference Series*. Vol. 1255. No. 1. IOP Publishing, 2019.

42. Dey, Ayon. “Machine Learning Algorithms : A Review.” (2016).

43. Fayyad, Usama M., et al., eds. "Advances in knowledge discovery and data mining." American Association for Artificial Intelligence, 1996.

44. Fletcher, Graham P., and Chris J. Hinde. "Interpretation of neural networks as Boolean transfer functions." *Knowledge-Based Systems* 7.3 (1994): 207-214.

45. Fu, King-Sun, and Huang Teddy. "Stochastic grammars and languages." International Journal of Computer & Information Sciences 1.2 (1972): 135-170.

46. Gamon, Michael. "Linguistic correlates of style: authorship classification with deep linguistic analysis features." *Coling 2004: Proceedings of the 20th international conference on computational linguistics*. 2004. doi: 10.3115/1220355.1220443.

47. Gavankar, Sachin S., and Sudhirkumar D. Sawarkar. "Eager decision tree." *2017 2nd International Conference for Convergence in Technology (I2CT)*. IEEE, 2017, DOI: https://doi.org/10.1109/I2CT.2017.8226246.

48. Golub, T. V., and Tyagunova M. Yu. "Method of steaming Ukrainian-language texts for classification of documents based on Porter's algorithm." *Scientific works of Donetsk National Technical University. Series: Informatics, cybernetics and computer engineering* 1.2017 (2017): 59-63.

49. Gómez-Adorno, Helena, et al. "A graph based authorship identification approach." *Working notes papers of the CLEF* (2015).

50. Gómez-Adorno, Helena, et al. "Document embeddings learned on various types of n-grams for cross-topic authorship attribution." *Computing* 100 (2018): 741-756.

51. “Great electronic dictionary of Ukrainian language (VESUM).” https://github.com/brown-uk/dict_uk.

52. Gupta, Gaurav. "A self explanatory review of decision tree classifiers." *International conference on recent advances and innovations in engineering (ICRAIE-2014)*. IEEE, 2014.

53. Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications." *Journal of emerging technologies in web intelligence* 1.1 (2009): 60-76.

54. Hardcastle, R. A. "CUSUM: a credible method for the determination of authorship?." *Science & Justice: Journal of the Forensic Science Society* 37.2 (1997): 129-138. doi: 10.1016/s1355-0306(97)72158-0.

55. Hasler, Eva, et al. "A comparison of neural models for word ordering." *arXiv preprint arXiv:1708.01809* (2017). doi:10.48550/arXiv.1708.01809.

56. Hearst, M. A. *Text data mining: Issues, techniques, and the relationship to information access*. Presentation notes for UW/MS workshop on data mining, 1997.

57. Hoover, David L. "Frequent word sequences and statistical stylistics." *Literary and Linguistic Computing* 17.2 (2002): 157-180. doi: 10.1093/llc/17.2.157.

58. Houvardas, John, and Efstathios Stamatatos. "N-gram feature selection for authorship identification." *International conference on artificial intelligence: Methodology, systems, and applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

59. Iyer, Rahul Radhakrishnan, and Carolyn Penstein Rose. "A machine learning framework for authorship identification from texts." *arXiv preprint arXiv:1912.10204* (2019).

60. Jafariakinabad, Fereshteh, and Kien A. Hua. "A self-supervised representation learning of sentence structure for authorship attribution." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.4 (2022): 1-16. doi: 10.1145/3491203.

61. Ji, Yangfeng, and Eisenstein, Jacob. "Representation learning for text-level discourse parsing." *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2014, pp. 13–24.

62. Juola, Patrick. "Authorship attribution." *Foundations and Trends® in Information Retrieval* 1.3 (2008): 233-334. doi:10.1561/ 1500000005.

63. Jusoh, Shaidah and Hejab, Al Fawareh. “Natural language interface for online sales systems”. *International Conference on Intelligent and Advanced Systems*, ICIAS 2007. 224 - 228. doi: 10.1109/ICIAS.2007.4658379.

64. Khomytska, I., Bazylevych, I., Teslyuk, V. "The Statistical Parameters of Ivan Franko's Authorial Style Determined by the Chi-square Test. " *In 17th IEEE International Conference on Computer Science and Information Technologies* (2022): 73–76. doi: 10.1109/CSIT56902.2022.10000491.

65. Kim, Hyunsoo, et al. "Dimension reduction in text classification with support vector machines." *Journal of machine learning research* 6.1 (2005).

66. Kohan, Ya O. "On the possibilities of formalizing natural languages." (2016): 137-143.

67. Koppel, Moshe, Schler Jonathan and Argamon Shlomo. "Computational methods in authorship attribution." *Journal of the American Society for information Science and Technology* 60.1 (2009): 9-26. doi: 10.1002/asi.20961.

68. Koppel, Moshe, Schler Jonathan, and Argamon Shlomo. "Authorship attribution: What's easy and what's hard." *JL & Pol'y* 21 (2012): 317. pp 282-289. doi: 10.1007/978-3-642-32790-2_34.

69. Köhler, R., Altmann, G. “Aims and Methods of Quantitative Linguistics.” *Problems of Quantitative.* Linguistics Chernivci. 2005. pp. 12-42.

70. Kruczek, Jakub, Kruczek, Paulina, and Kut, Marcin. "Are n-gram categories helpful in text classification?." *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20*. Springer International Publishing, 2020. DOI: https://doi.org/10.1007/978-3-030-50417-5_39.

71. Kuropiatnyk, Olena S., and Shynkarenko, Viktor I.. "Automation of template formation to identify the structure of natural language documents." *5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, 2021.

72. Priyanka, and Kumar, Dharmender. "Decision tree classifier: a detailed survey." *International Journal of Information and Decision Sciences* 12.3 (2020): 246-269.

73. Langseth, Helge, and Nielsen, Thomas D. "Classification using hierarchical naive Bayes models." *Machine learning* 63 (2006): 135-159.

74. Li, Jiaqi, et al. "A survey of discourse parsing." *Frontiers of Computer Science* 16.5 (2022): 165329.

75. Liu, Pengfei, et al. "Contextualized non-local neural networks for sequence learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019. doi: 10.1609/aaai.v33i01.33016762.

76. Liu, Qiao, et al. "Content attention model for aspect based sentiment analysis." *Proceedings of the 2018 world wide web conference*. 2018. 1023–1032. doi: 10.1145/3178876.3186001.

77. Liu, Zhiyuan, et al. "Sentence representation." *Representation Learning for Natural Language Processing* (2020): 59-89. doi:10.1007/978-981-15-5573-2_4.

78. Luo, Xiaoyu. "Efficient English text classification using selected machine learning techniques." *Alexandria Engineering Journal* 60.3 (2021): 3401-3409.

79. Lupei, Maksym, et al. "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks." (2020): 30-36. doi: 10.15587/1729-4061.2020.195041.

80. Luyckx, Kim, and Daelemans, Walter. "The effect of author set size and data size in authorship attribution." *Literary and linguistic Computing* 26.1 (2011): 35-55.

81. Lytvyn, Vasyl, et al. "Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology." *Mathematics* 11.4 (2023): 904.

82. Lytvyn, Vasyl, et al. "Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution." *Eastern-European Journal of Enterprise Technologies* 6.2 (2019): 28-51. doi:10.15587/1729-4061.2019.186834.

83. Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR). [Internet]* 9.1 (2020): 381-386.

84. Markov, Ilia, Baptista, Jorge, and Pichardo-Lagunas, Obdulia. "Authorship attribution in portuguese using character n-grams." *Acta Polytechnica Hungarica* 14.3 (2017): 59-78. doi: 10.12700/APH.14.3.2017.3.4.

85. Marwan, Norbert, et al. "Recurrence plots for the analysis of complex systems." *Physics reports* 438.5-6 (2007): 237-329. doi: 10.1016/j.physrep.2006.11.001.

86. Marwan N. How to avoid potential pitfalls in recurrence plot based data analysis. International Journal of Bifurcation and Chaos 21 (4) (2011) 1003–1017. doi: 10.1142/S0218127411029008

87. Matthew E Peters, Neumann Mark, Iyyer Mohit, Gardner Matt, Clark Christopher, Lee Kenton, and Zettlemoyer Luke. 2018. Deep contextualized word representations. arXiv. doi: 10.48550/arXiv.1802.05365.

88. Mazzei, Alessandro, and Lombardo Vincenzo. "Building a Large Grammar for Italian." *LREC*. 2004.

89. Mohammad, AL-Smadi, et al. "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features." *Information Processing & Management* 53.3 (2017): 640-652. doi:10.1016/j.ipm.2017.01.002.

90. Moshkina, Vadim, Ilya Andreeva, and Nadezhda Yarushkinaa. "Solving the problem of determining the author of text data using a combined assessment." *CEUR Workshop Proceedings*. Vol. 2782. 2020.112-118.

91. Mrva, Jakub, et al. "Decision support in medical data using 3D decision tree visualisation." *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, 2019. DOI: https://doi.org/10.1109/EHB47216.2019.8969926.

92. Nykonenko, Andrii, et al. “About authorship attribution system”. *Artificial Intelligence* 2 (2016): 77-85.

93. Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).

94. Pokou, Yao Marc Jean, Fournier-Viger Philippe, and Moghrabi Chadia. "Authorship Attribution Using Small Sets of Frequent Part-of-Speech Skip-grams." *Flairs conference*. 2016.

95. Popescu, Ioan-Iovitz, and Altmann Gabriel. "Some aspects of word frequencies." *Glottometrics* 13 (2006): 23-46.

96. Popescu, Ioan-Iovitz. *Word frequency studies*. Mouton de Gruyter, 2009.

97. Raheja, J. L., Mishra Anand, and Chaudhary Ankit. "Indian sign language recognition using SVM." *Pattern Recognition and Image Analysis* 26 (2016): 434-441.

98. Rana, Soraya. *Examining the role of local optima and schema processing in genetic search*. Diss. Colorado State University, 1999.

99. Russell, Stuart J., and Norvig Peter. *Artificial intelligence a modern approach*. London, 2010.

100. Rygl, Jan, and Horák Aleš. "Authorship Attribution: Comparison of Single-Layer and Double-Layer Machine Learning." *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings 15*. Springer Berlin Heidelberg, 2012. pp 282-289. doi: 10.1007/978-3-642-32790-2_34.

101. Sablin, Oleg, et al. "Rational distribution of excess regenerative energy in electric transport systems on the basis of fuzzy logic application." *Archives of Transport* (2017).

102. Segaran, T. *Programming Collective Intelligence*. O'Reilly Media Inc, 2007.

103. Shieber, Stuart M. "Evidence against the context-freeness of natural language." *The Formal complexity of natural language*. Dordrecht: Springer Netherlands, 1985. 320-334.

104. Shynkarenko, Viktor, et al. "Modeling of lightning flashes in thunderstorm front by constructive production of fractal time series." *Conference on Computer Science and Information Technologies*. Cham: Springer International Publishing, 2019.

105. Shynkarenko, V. I. "Constructive-Synthesizing Representation of Geometric Fractals". *Cybernetics and Systems Analysis*. 55.2 (2019): 189-199.

106. Shynkarenko, V., and Zhuchyi, L. "Constructive-synthesizing modeling of ontological document management support for the railway train speed restrictions". *Science and Transport Progress* 2.98 (2022): 59-68. doi: 10.15802/stp2022/268001.

107 Shynkarenko, Viktor, and Demidovich Inna. "Constructive-synthesizing modeling of natural language texts." *Computer systems and information technologies* 32023, p. 81-91.

108. Shynkarenko, V. I., and Demidovich I. M.."Determination of the attributes of authorship of natural texts." *Artificial intelligence* 3 (2018): 27-35.

109. Shynkarenko, Viktor, and Demidovich Inna. "Natural Language Texts Authorship Establishing Based on the Sentences Structure." *COLINS*, 2022, p. 328-337.

110. Shynkarenko, V. I., Demidovich, I. M., and Kuropiatnyk, O. S. "A Dual Approach to Establishing the Authority of Technical Natural Language Texts and Their Components." *Science and Transport Progress* 2 (102) (2023): 71-85. doi: 10.15802/stp2023/288958.

111. Shynkarenko, V. I., and Demydovych I. M. "Methods and software for significant indicators determination of the natural language texts author profile." *Problems in programming* 3 (2023): 22-29. doi: 10.15407/pp2023.03.22

112. Shynkarenko, Viktor I., et al. " Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task." *IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT),* 2021, p. 48-51. doi: 10.1109/CSIT52700.2021.9648829.

113. Shynkarenko, Viktor I., and Demidovich Inna. "Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights." *5th International Conference on Computational Linguistics and Intelligent Systems (COLINS),* 2021, p. 832-844.

114. Shynkarenko, V. I., Ilchenko P. V., and Zabula H. V. "Tools of investigation of time and functional efficiency of bionic algorithms for function optimization problems." *Problems in programming* 2-3 (2018): 270-279.

115. Shynkarenko, V. I., and Ilman V. M. "Constructive-synthesizing structures and their grammatical interpretations. I. Generalized formal constructive-synthesizing structure." *Cybernetics and Systems Analysis* 50.5 (2014): 655-662.

116. Shynkarenko, V. I., and Kuropiatnyk O. S. "Constructive-synthesizing model of text graph representation." *PROBLEMS IN PROGRAMMING* 2-3 (2016): 63-72.

117. Shynkarenko, Viktor I., and Olena S. Kuropiatnyk. "Constructive model of the natural language." *Acta Cybernetica*, 23.4 (2018): 995-1015.

118. Shynkarenko, V., and Zhevaho, O. "Constructive Modeling of the Software Development Process for Modern Code Review. " *International Scientific and Technical Conference on Computer Sciences and Information Technologies* 1 (2020): 392–395.

119. Shynkarenko, V., and Zhevaho, O. "Development of a toolkit for analyzing software debugging processes using the constructive approach." *Eastern-European Journal of Enterprise Technologies*. 5/2.107 (2020): 29-38.

120. Silberztein, Max. "A new linguistic engine for nooj: Parsing context-sensitive grammars with finite-state machines." *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications: 11th International Conference, NooJ 2017, Kenitra and Rabat, Morocco, May 18–20, 2017, Revised Selected Papers 11*. Springer International Publishing, 2018. p. 240– 250.

121. Srinivas, Ranjini. "Managing Large Data Sets Using Support Vector Machines." 2010.

122. Sidorov, Grigori O. "Automatic authorship attribution using syllables as classification features." *Rhema. Рема* 1 (2018): 62-81.

123. Sisodia, Pushpendra Singh, Tiwari Vivekanand, and Kumar Anil. "A comparative analysis of remote sensing image classification techniques." *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2014. doi: 10.1109/ICACCI.2014.6968245.

124. Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology* 60.3 (2009): 538-556.

125. Stamatatos, Ph D. "On the robustness of authorship attribution based on character n-gram features." *Journal of Law and Policy* 21.2 (2013): 427–439.

126. Szwed, Piotr. "Authorship attribution for polish texts based on part of speech tagging." *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation: 13th International Conference, BDAS 2017, Ustroń, Poland, May 30-June 2, 2017, Proceedings 13*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-58274-0_26.

127. Tal, Benjamin. "Neural Network-Based System of Leading Indicators." *CIBC World markets*, 2003.

128. Tenney, Ian, et al. "What do you learn from context? probing for sentence structure in contextualized word representations." *arXiv preprint arXiv:1905.06316* (2019). doi: 10.48550/arXiv.1905.06316.

129. Tou, J.T., and Gonzalez, R.C. “Pattern Recognition Principles.” *Addison-Wesley Publishing Company*. 1974. p. 377.

130. Towell, Geoffrey G., and Shavlik Jude W. "Extracting refined rules from knowledge-based neural networks." *Machine learning* 13 (1993): 71-101.

131. Tu, Jack V. "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes." *Journal of clinical epidemiology* 49.11 (1996): 1225-1231.

132. Vapnik, Vladimir. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

133. Vapnik, Vladimir. *The nature of statistical learning theory*. Springer science & business media, 1999.

134. Vijayarani, S., and Muthulakshmi M. "Comparative analysis of bayes and lazy classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 2.8 (2013): 3118-3124.

135. Vijayarani, S., Ilamathi J. and Nithya. "Preprocessing techniques for text mining-an overview." *International Journal of Computer Science & Communication Networks* 5.1 (2015): 7-16.

136. Vysotska, Victoria, Holoshchuk Svitlana, and Holoshchuk Roman. "A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach." *COLINS*. 2021.

137. Vysotska, Victoria, et al. "Method of similar textual content selection based on thematic information retrieval." *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*. Vol. 3. IEEE, 2019.

138. Walenski, Matthew, et al. "Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies." *Human brain mapping* 40.8 (2019): 2275-2304. doi: 10.1002/hbm.24523.

139. Wang, Li-Min, et al. "Combining decision tree and Naive Bayes for classification." *Knowledge-Based Systems* 19.7 (2006): 511-515.

140. Wang, Wei, et al. "Structbert: Incorporating language structures into pre-training for deep language understanding." *arXiv preprint arXiv:1908.04577* (2019). doi:10.48550/arXiv.1908.04577.

141. White, Lyndon, et al. "Sentence Representations and Beyond." *Neural representations of natural language* (2019): 93-114. doi: 10.1007/978-981-13-0062-2_5.

142. Whitley, Darrell. "A genetic algorithm tutorial." *Statistics and computing* 4, 1994, 65-85.

143. Xhemali, Daniela, Hinde Chris J. and Stone Roger. "Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009).

144. Yalcin, Kadir, Cicekli Ilyas, and Ercan Gonenc. "An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding." *Expert Systems with Applications* 197 (2022): 116677.

145. Yang, Feng-Jen. "An extended idea about decision trees." *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2019. doi: 10.1109/CSCI49370.2019.00068.

146. Yunita Sari, Vlachos Andreas, Stevenson Mark, Continuous N-gram Representations for Authorship Attribution, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp.267-273. doi: 10.18653/v1/E17-2043.

147. Zeldes, Amir, and Schroeder Caroline T. "Computational methods for coptic: Developing and using part-of-speech tagging for digital scholarship in the humanities." *Digital Scholarship in the Humanities* 30.suppl_1 (2015): i164-i176. doi: 10.1093/llc/fqv043.

148. Zbilut, Joseph P., and Charles L. Webber Jr. "Embeddings and delays as derived from quantification of recurrence plots." *Physics letters A* 171.3-4 (1992): 199-203.

ДОДАТОК А

Акт впровадження

"ЗАТВЕРДЖУЮ"
Перший проректор
Українського державного університету
науки і технологій
_____ д.т.н., проф. Анатолій РАДКЕВИЧ
11.12 _____ 2023 р.

АКТ
Про використання результатів дисертації

"РОЗВИТОК МЕТОДІВ ТА ЗАСОБІВ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНИХ ТЕКСТІВ НА ОСНОВІ КОНСТРУКТИВНО-ПРОДУКЦІЙНОГО МОДЕЛЮВАННЯ"
Демидович Інни Миколаївни,
Представленої на здобуття наукового ступеня доктора філософії
спеціальності 122 «Комп'ютерні науки»

Цей акт складений про те, що у навчальному процесі Дніпропетровського інституту інфраструктури і транспорту Українського державного університету науки і технологій при підготовці аспірантів за спеціальністю 122 «Комп'ютерні науки» використовуються наукові та практичні результати, що отримані в дисертації Демидович І. М., при викладенні дисципліни «Ефективність інформаційних систем та комп'ютерних технологій».

Зав. кафедри
Комп'ютерні інформаційні технології
к.т.н., доцент Вадим ГОРЯЧКІН

Декан факультету
Комп'ютерні технології і системи
к.т.н, доцент Володимир МАЛОВІЧКО