



Ідентифікація автора: дослідження авторського стилю написання тексту

ГЛАВЧЕВА ЮЛІЯ МИКОЛАЇВНА,
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»,
GLAVCHEVA@KHPI.EDU.UA

Авторський стиль



Ідентифікація автора – визначення потенційного автора документа на основі вимірювання деяких текстових функцій (характеристик тексту).

Авторський стиль - власний письмовий стиль автора, який не підсвідомо використовується автором при написанні текстів.

Лінгвометрія (стилометрія)

Дослідження явища авторства:

- *яка різниця між стилями авторів?*
- *що робить мову автора динамічною, легкою для сприйняття?*
- *які характеристики є індивідуальними?*
- *які характеристики можна вважати загальним?*

Стилометрія спирається на те, що кожен автор має незамінні письмові особливості, які не можна імітувати.

Перший етап досліджень

(до кінця 1990-х р.)

-
- Текстові дані є **завеликими** (наприклад, книги) і не є стилістично однорідними.
 - Кількість потенційних авторів-кандидатів замала (зазвичай 2 або 3).
 - Досліджувані тексти не класифікуються за темою.
 - Відсутність коректних методів оцінки результатів досліджень.

Оцінка запропонованих методів була в основному інтуїтивно зрозумілою (зазвичай на основі суб'єктивного візуального перегляду).

- Складність порівняння різних методів з причини відсутності необхідних даних для дослідження.

Другий етап досліджень

(з початку 2000-х р.)

Стрімкий розвиток досліджень у наукових сферах: **пошук інформації, машинне навчання та обробка природної мови (НЛП)** вплинув на технологію ідентифікації автора:

- розроблено ефективні методи для представлення та класифікації текстів великих обсягів;
- започатковано використання алгоритмів машинного навчання для обробки багатовимірних даних;
- розроблено стандартні методології оцінки для порівнювати різних підходів;
- розроблено інструменти **обробки природної мови (NLP – Natural Language Processing)**, здатні ефективно аналізувати текст та надавати нові характеристики для представлення стилю (наприклад, на основі синтаксичних функцій).

Сучасний етап

(з кінця 2000 р)

Розробка практичних додатків, що стосуються текстів **реального світу** (наприклад, електронна пошта, **академічні тексти, наукові публікації**, блоги, повідомлення в Інтернеті, вихідний код і т. д.), а не вирішення суперечливих літературних питань.

Удосконалення:

- об'єктивної оцінки методів, порівняння різних методів, заснованих на спільних контрольних даних;
- розгляд факторів, що відіграють важливу роль у точності розроблених моделей (наприклад, розмір навчального тексту, кількість авторів кандидатів).

Сфери застосування

- розвідка (наприклад, присвоєння повідомлень відомим терористам);
- кримінальне право (наприклад, виявлення авторів образливих повідомлень);
- цивільне право (наприклад, авторські спори);
- комп'ютерна криміналістична експертиза (наприклад, ідентифікація авторів вихідного коду шкідливого програмного забезпечення);
- літературні дослідження (наприклад, приписування анонімних літературних творів відомим авторам).

Завдання аналізу авторства

- перевірка авторства (чи був даний текст написаний певним автором чи ні);
- **виявлення плагіату** (знаходження подібності між двома текстами);
- визначення характеристик авторів (інформація про вік, освіту, стать автора даного тексту);
- виявлення стилістичних невідповідностей (пошук відмінностей у написанні спільної роботи).

Виявлення плагіату



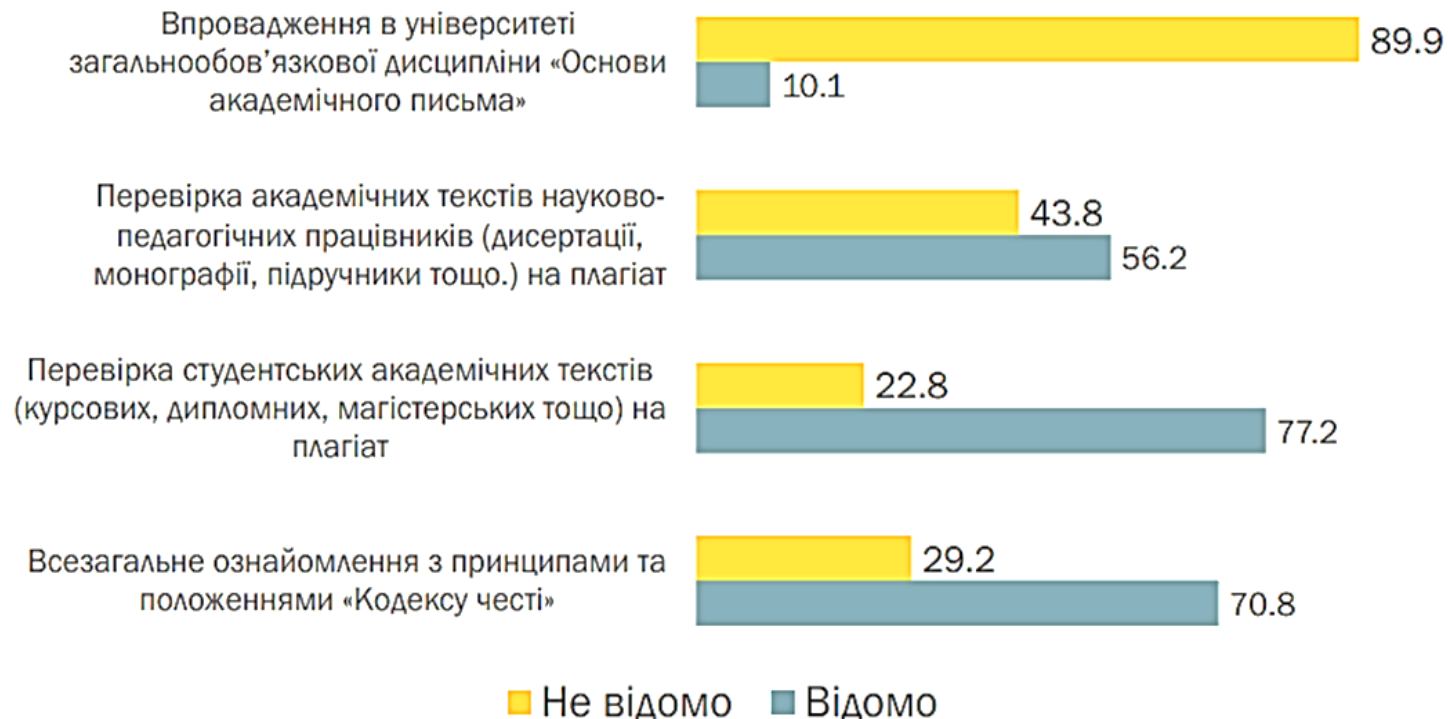
Причини

1. Велика кількість та **доступність електронної інформації** (у тому числі сумнівної якості).
2. Наявність інформаційних інструментів для швидкої обробки текстів.
3. Низький **рівень академічної доброчесності** (*АНТИКОРУПЦІЙНА ПРОГРАМА Міністерства освіти і науки України на 2018 рік*).
4. Підходи до **оцінки наукової діяльності** вчених та університетів, що використовуються.
5. **Відсутність загальної бази даних** наукових текстів українською мовою (*АНТИКОРУПЦІЙНА ПРОГРАМА Міністерства освіти і науки України на 2018 рік*).
6. **Неузгодженість вимог** до закладів вищої освіти, науково-педагогічних працівників та здобувачів освіти з **фінансуванням та іншим ресурсним забезпеченням вищої освіти і науки** (Методичні рекомендації для закладів вищої освіти з підтримки принципів академічної доброчесності, МОН, №1/9-650 від 23.10.2018)

Онлайн-опитування проводилось за анкетною складеною фахівцями
НДЦ ПС «Соціоплюс» КПІ ім. Ігоря Сікорського за допомогою сервісу Google Forms.

http://ela.kpi.ua/bitstream/123456789/23076/1/Akademichna_dobrochesnist.pdf

Наскільки Вам відомо про наступні складові політики
КПІ ім. Ігоря Сікорського в дотриманні принципів та правил
академічної доброчесності?



Всього у опитуванні взяли участь
907 респондентів.

Тривалість опитування: **лютий-травень 2018 року.**

Наскільки поширені серед Ваших однокурсників/колеґ наступні прояви поведінки

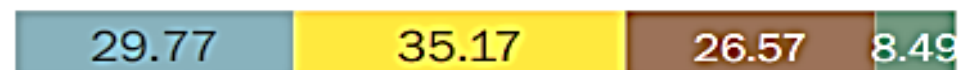
самоплаґіат – оприлюднення (частково або повністю)
власних раніше опублікованих наукових результатів як
нових



компіляція – створення значного масиву тексту без
поглибленого вивчення проблеми шляхом копіювання
тексту



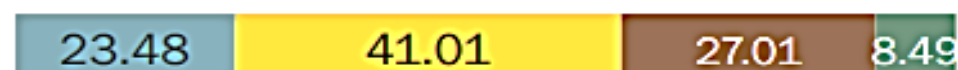
парафраза – переказ своїми словами чужих думок, ідей
або тексту; сутність парафрази полягає в заміні слів
(знаків)



представлення суміші власних і запозичених аргументів
без належного цитування

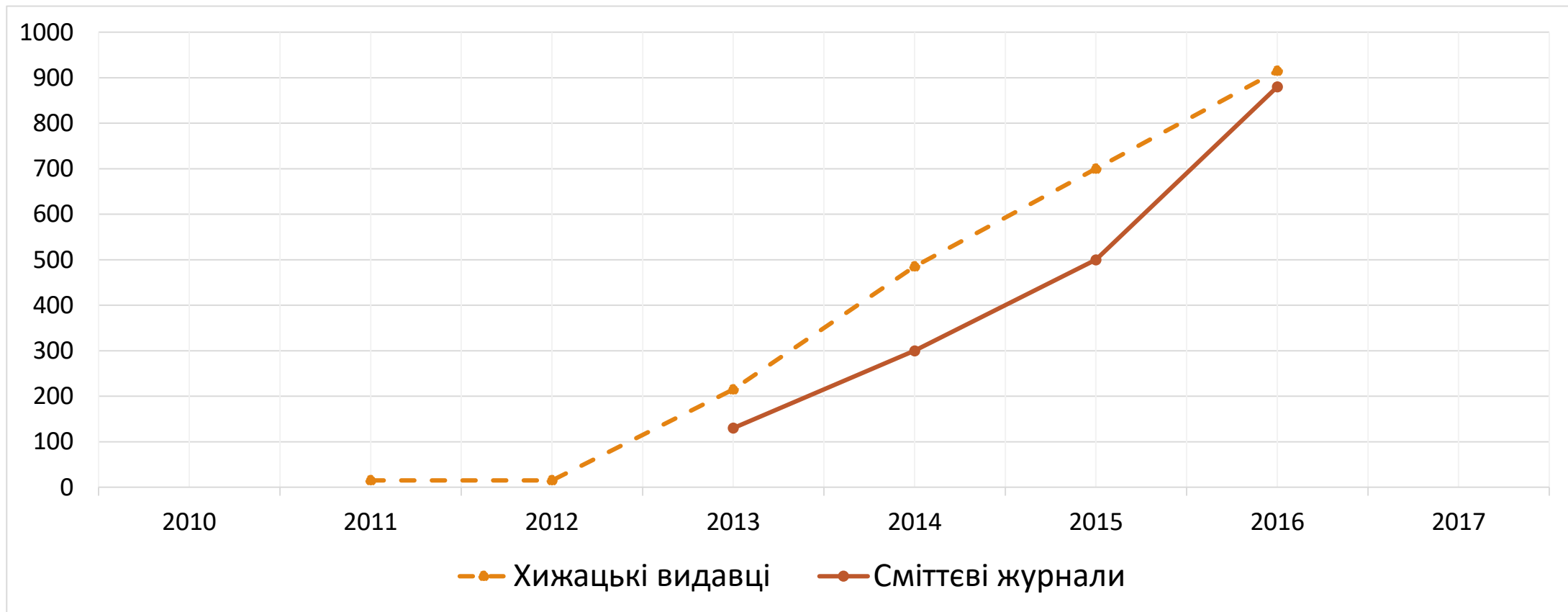


внесення незначних правок у скопійований матеріал
(перефразування речень, зміна порядку слів у них тощо)
та без належного оформлення цитування



■ Дуже поширено ■ Швидше поширено ■ Швидше не поширено ■ Взагалі не поширено

Динаміка зростання кількості «хижацьких» видавців і «сміттєвих» журналів



Розподіл видів плагіату на умовні групи

Текстовий плагіат		Плагіат ідей
Принцип роботи: пошук подібностей у тексті	Принцип роботи: пошук стилістичних особливостей автора при написанні тексту	Завдання вирішується науковцями
Правовий аспект: зазвичай система завантажує тексти, що перевіряє.	Правовий аспект: Система зберігає показники, що характеризують авторський стиль написання роботи	
Необхідні умови для ефективної роботи: наявність великої бази даних електронних текстів певною мовою для проведення пошуку подібностей	Необхідні умови для ефективної роботи: наявність оригінального документа певного автора	
Обмеження: документ у цифровій формі відсутній.	Обмеження: відсутність оригінального документа певного автора.	

Авторський стиль написання тексту

Основні стилістичні ознаки, за якими проводяться дослідження письмового стилю, ідентифікують на таких етапах обробки природньої мови:

1. Лексичний аналіз
2. Символьний аналіз
3. Синтаксичний аналіз
4. Семантичний аналіз



Експерименти

Метою роботи є виявлення стилістичних особливостей авторів для визначення авторського стиля та запозичень в текстах українською мовою

Корпус текстів 1: 437 одноосібних наукових статей 38 науковців України

Корпус текстів 2: 108 одноосібних наукових статей (розмічені) 10 науковців України

Джерело даних: репозитарій Національного технічного університету "Харківський політехнічний інститут" <http://repository.kpi.kharkov.ua>

Корпус текстів (структурний підхід)

<author> Автор**</author>**

<title> Назва**</title>**

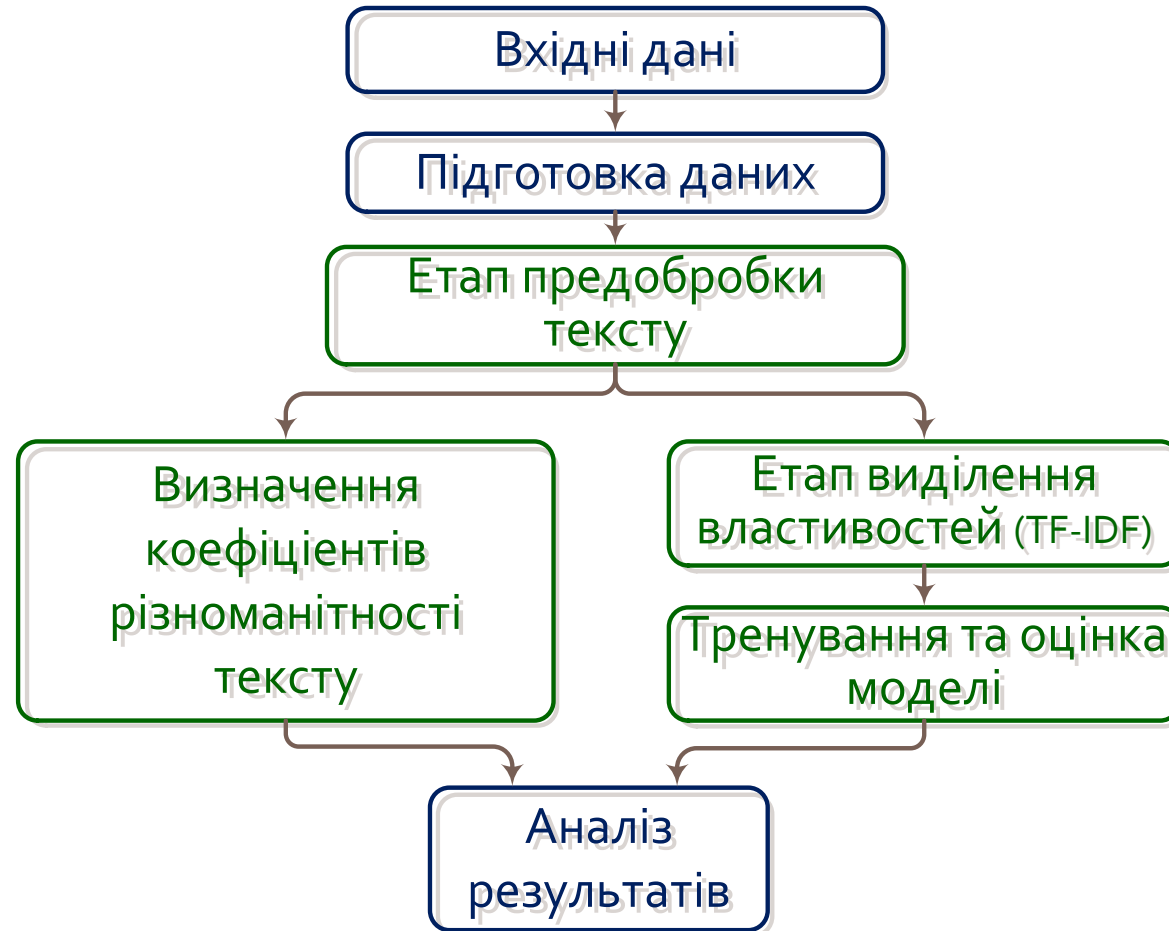
<description>Анотація, ключові слова українською**</description>**

<text> Основний текст статті**</text>**

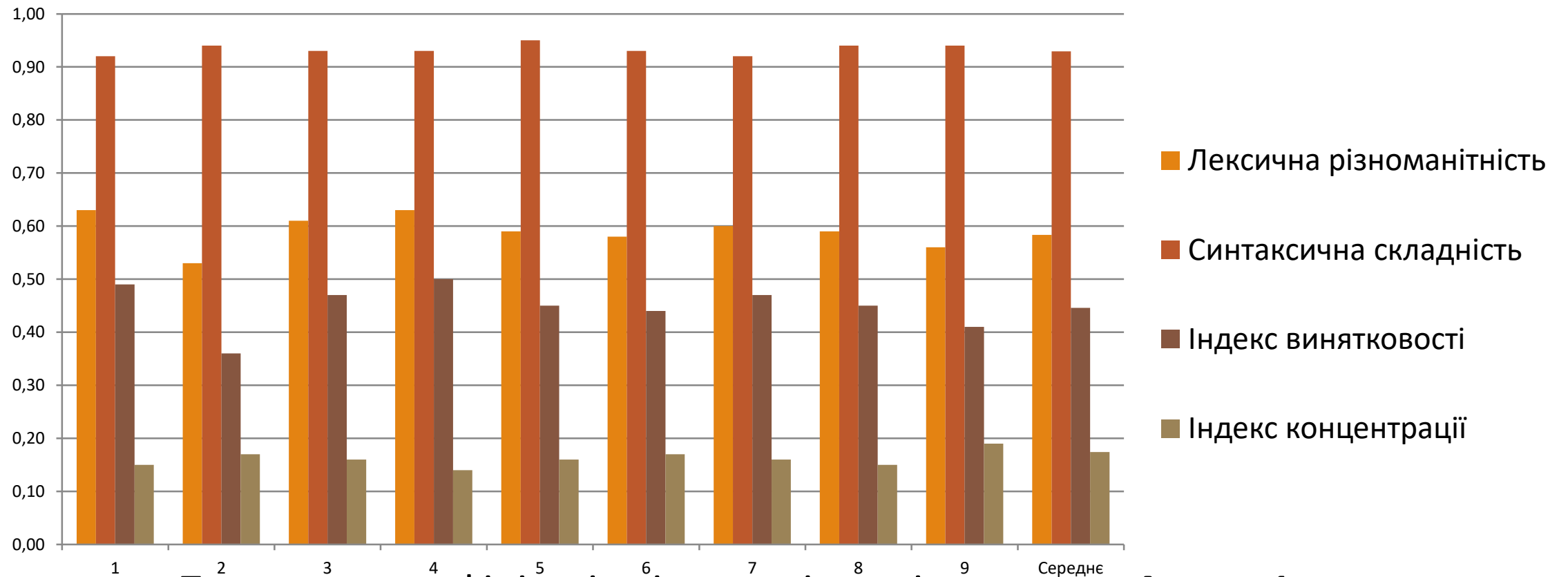
<references> Перелік джерел**</references>**

<inf> Інші дані**</inf>**

Загальна схема роботи для визначення стилю автора

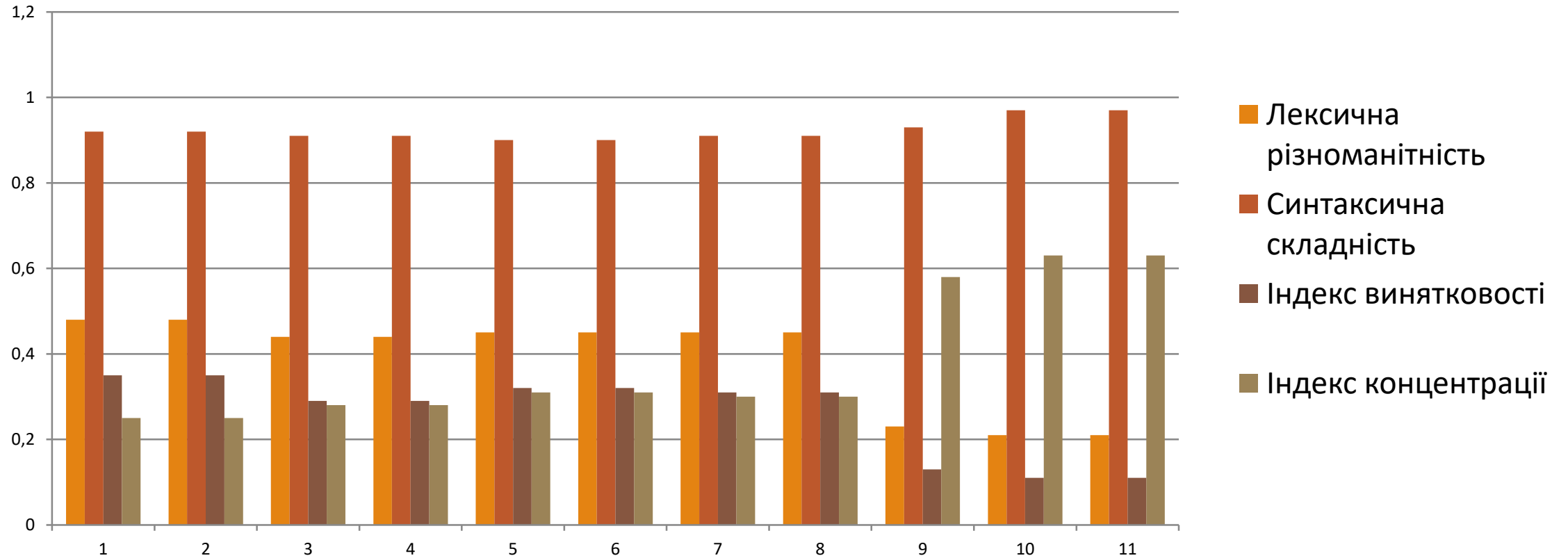


Результати (1)



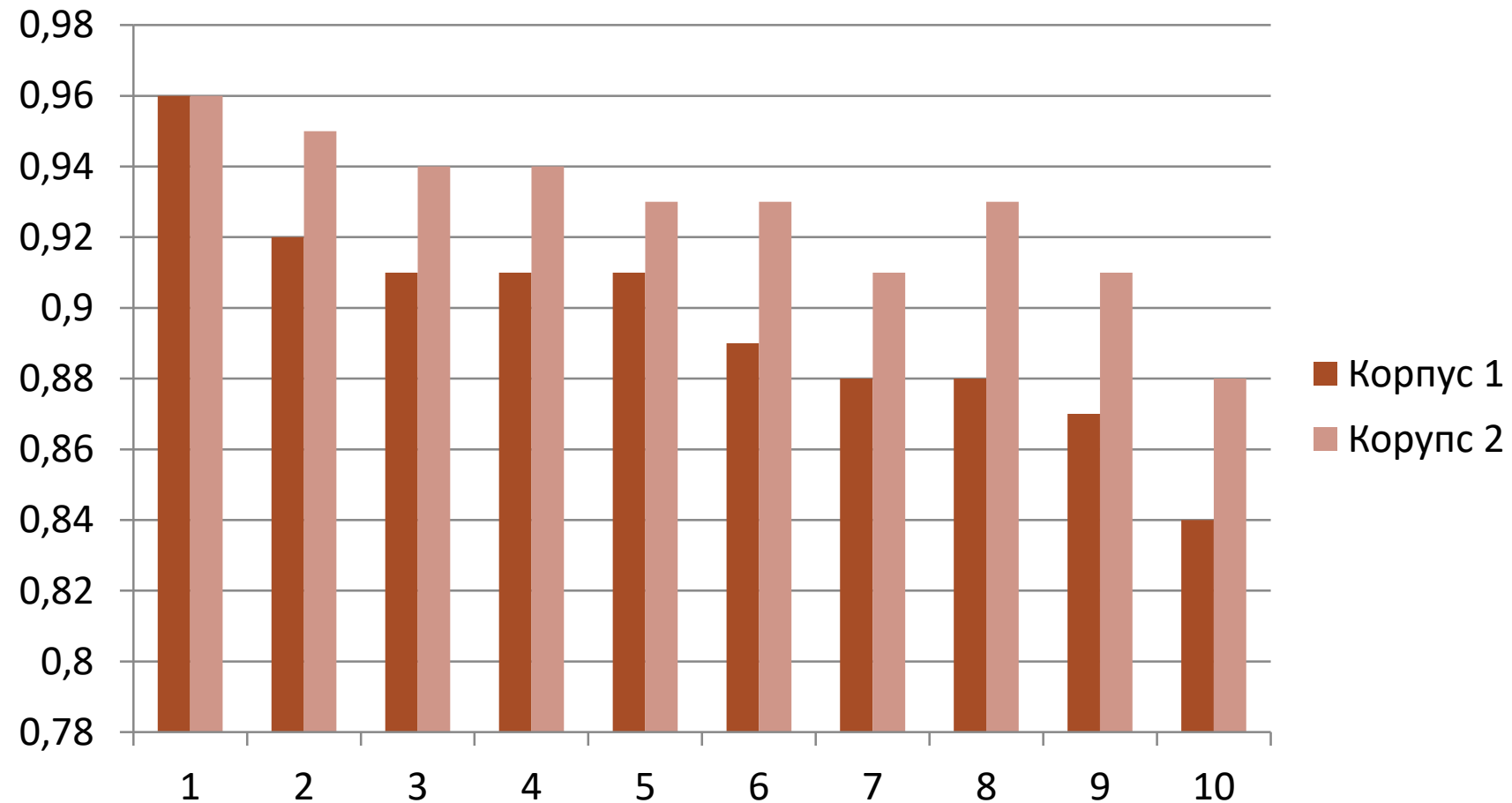
Показники коефіцієнтів різноманітності тексту для **Автора1**

Результати (2)

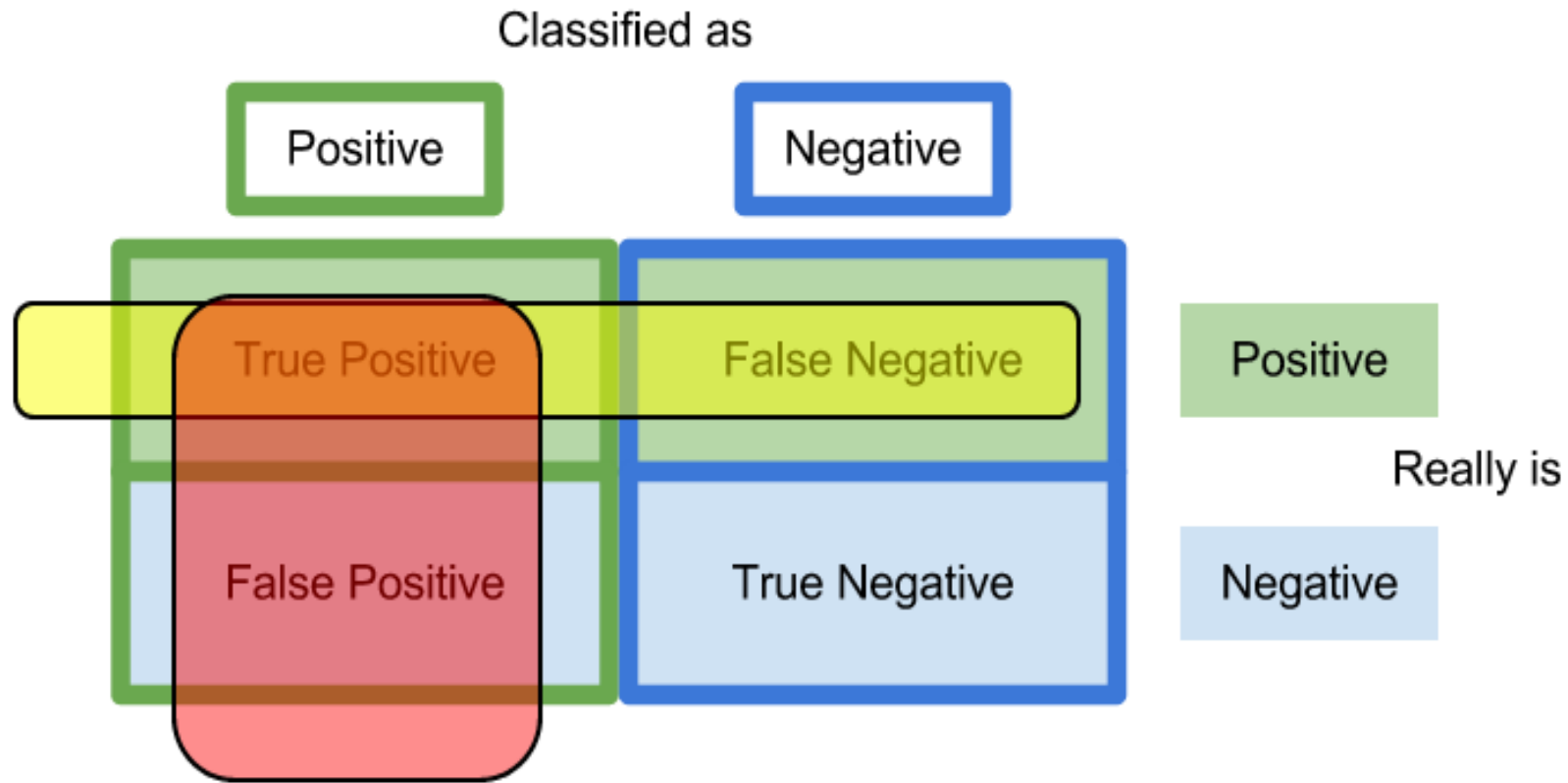


Показники коефіцієнтів різноманітності тексту для **Автора2**

Коефіцієнт синтаксичної складності (Автор 3)



Міра оцінки якості



РЕЗУЛЬТАТИ першого експерименту (МЕТОД БАЕСА) для *Корпусу текстів 1*

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Precision} = 0.92$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Recall} = 0.72$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{F-measure} = 0.83$$

Визначення авторського стилю написання роботи



<https://www.digitaltrends.com/cool-tech/emma-identity-ai-web-app/>

Emma Identity для англійської (1)

EMMA

Press Kit

Contact creators

Tutorial

Sign in

Emma requires 5,000 words

Emma needs only 5,000 words to learn each unique writing style.

More than 50 math parameters stand behind every author's writing identity. As thousands of tests on works of novelists, journalists, researchers, professors and students show, Emma proves to be 85% accurate in attributing authorship.

This is the best result in the authorship attribution ever achieved by engineers worldwide.



85% Accuracy Defining Authorship

<https://emmaidentity.com>

ВИСНОВКИ

- розглянуто завдання ідентифікації автору текстового документу
- підтверджено актуальність задачі визначення стилю автора для виявлення можливого плагіату в текстах українською мовою
- досліджено вплив якості текстового корпусу на точність результатів
- підтверджено, що створення текстового корпусу для визначення стилю автора є невід'ємним завданням разом з розробкою технологій та методів визначення стилю автора
- розглянуто підхід для визначення стилю автора на прикладі розроблено корпусу українських академічних робіт
- виділено перспективні напрямки розвитку для визначення плагіату та встановлення авторства документу

Література:

1. Stamatatos E. A survey of modern authorship attribution methods //Journal of the American Society for information Science and Technology. – 2009. – Т. 60. – №. 3. – С. 538-556.
2. Каніщева О. В. Визначення стилю автора для виявлення плагіату в академічному середовищі / О. В. Каніщева, Ю. Н. Главчева, В. А. Висоцька // Системний аналіз та інформаційні технології : матеріали 19-ї Міжнар. наук.-технічн. конф. SAIT 2017, 22-25 травня 2017. — Київ : ННК "ІПСА" НТУУ "КПІ" ім. Ігоря Сікорського, 2017. — С. 78-79.
3. Vysotska V., Kanishcheva O., Hlavcheva Y. Authorship Identification of the Scientific Text in Ukrainian with Using the Lingvometry Methods //2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). – IEEE, 2018. – Т. 2. – С. 34-38.



Дякую!

ГЛАВЧЕВА ЮЛІЯ МИКОЛАЇВНА,
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»,
GLAVSCHEVA@KHPI.EDU.UA