Розглянуто особливості застосування технологій лінгвометрії, стилеметрії та глоттохронології для визначення стилю автора публікацій. Лінгвостатистичний аналіз авторського тексту використовує переваги контент-моніторінгу на основі методів NLP для визначення стопових слів. Квантитативний аналіз стопових слів використано для визначення степеня приналежності аналізованого тексту конкретному авторові. Запропоновано метод визначення стилю автора україномовного тексту

Ключові слова: стиль автора, статистичний лінгвістичний аналіз, квантитативна лінгвістика, авторська атрибуція

Рассмотрены особенности применения технологий лингвометрии, стилеметрии и глоттохронологии для определения стиля автора публикаций. Лингвостатистический анализ авторского текста использует преимущества контент-мониторинга на основе методов NLP для определения стоповых слов. Квантитативный анализ стоповых слов использовано для определения степени принадлежности анализируемого текста конкретному автору. Предложен метод определения стиля автора украиноязычного текста

Ключевые слова: стиль автора, статистический лингвистический анализ, квантытативная лингвистика, авторская атрибуция

УДК 004.89
DOI: 10.15587/1729-4061.2017.107512

РОЗРОБЛЕННЯ МЕТОДУ ВИЗНАЧЕННЯ СТИЛЮ АВТОРА УКРАЇНОМОВНИХ ТЕКСТІВ НА ОСНОВІ ТЕХНОЛОГІЙ ЛІНГВОМЕТРІЇ, СТИЛЕМЕТРІЇ ТА ГЛОТТОХРОНОЛОГІЇ

В. В. Литвин
Доктор технічних наук, професор*
E-mail: vasyl.v.lytvyn@lpnu.ua
В. А. Висоцька
Кандидат технічних наук, доцент*
E-mail: victoria.a.vysotska@lpnu.ua
П. Я. Пукач
Доктор технічних наук, доцент**
E-mail: petro.y.pukach@lpnu.ua
І. О. Бобик
Кандидат фізико-математичних наук, доцент**
E-mail: igor.bobyk@gmail.com
Д. І. Угрин
Кандидат технічних наук, доцент
Кафедра інформаційних систем
Чернівецький факультет Національного технічного університету «Харківський політехнічний інститут»
вул. Головна, 203-а, м. Чернівці, Україна, 58000
E-mail: ugrund38@gmail.com
*Кафедра інформаційних систем та мереж***
Кафедра вищої математики*
***Національний університет «Львівська політехніка»
вул. С. Бандери, 12, м. Львів, Україна, 79013

1. Вступ

Поштовхом статистичних лінгвістичних досліджень (квантитативна лінгвістика) стала поява та активний розвиток інформаційних технологій (ІТ) в напряму NLP та Web Mining [1]. На початку 1960-х років в Інституті мовознавства ім. О. О. Потебні АН УРСР організовано групу структурно-математичної лінгвістики [2]. Вона розпочала цілеспрямоване статистичне дослідження українських текстів художнього, науково-технічного та соціально-політичного функціональних стилів. Це дало змогу виявити їхні статистичні параметри. Тоді ж розпочався проект з укладання серії частотних словників: художньої прози, драми, поезії, публіцистики, наукової прози, до якого було також залучено лабораторію комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка (Україна) [3]. Основними напрямами прикладної статистичної лінгвістики та суміжних із нею наук є розроблення методів та технологій визначення статистичної структури тексту для розв'язування задач, зокрема, лінгвометрії [4], стилеметрії [5] та глотохронології [6]. Ці задачі полягають, наприклад, в автоматизації лексикографічних процесів, порівнянні словників, створенні систем стенографії, автоматичного визначення мови [7]. Для визначення стилю автора використовують для розв'язування лінгвістично статистичних задач:

– автоматичного визначення мови;

– розрахунку та аналізу коефіцієнтів лексичного авторського мовлення;

– визначення степеня плагіату;

– ідентифікації автора тексту або самого тексту;

– аналізу феномену авторства та динаміки зміни авторського стилю;

– визначення та аналізу степеня авторської атрибуції [8].

Важливими завданнями мовознавства є створення і порівняння словників за допомогою лінґвометрії (у тому числі частотних та статистичних), створення автоматичних словників, тезаурусів, створення систем стенографії, автоматичне визначення мови, інформаційний пошук тощо. Для моделювання деяких процесів контент-моніторингу та контент-аналізу знаходять статистичні і перехідні ймовірності морфем тексту. На основі побудованих таблиць моделюють перевірку досліджуваного слова на наявність помилки, пропонують декілька найбільш ймовірних варіантів.

Метою стилеметрії є типологія, атрибуція (авторська, часова, стильова – для застосування, наприклад, у судовій і кримінальній лінгвістиці), діагностика, реконструкція і т. ін. текстів та їх частин. Прикладом вирішення мовознавчої проблеми є процес авторської атрибуції уривків тексту. Для цього обчислюють частоти слововживань у аналізованих текстах. Використовуючи частотні словники творчості письменників загалом чи окремих їх творів, визначають автора твору (або твір – якщо це дозволяє словник).

Глоттохронологія досліджує швидкість мовних змін і визначає на цій основі час розділення споріднених мов і ступінь близькості між ними. Метод датування, що його застосовують для визначення тривалості роздільного існування двох споріднених мов, ґрунтується на припущенні про те, що основна частина лексичного складу будь-якої мови (ядерна лексика) змінюється з однаковою швидкістю і вимагає підрахунку процентного співвідношення спільних елементів у їхньому основному словнику.

Кожна мова має власні статистичні параметри, і знання частоти поя літер та їх сполучень (біграм, триграм, чотириграм) певної мови дає змогу автоматично її ідентифікувати. Наприклад, для українських текстів виявлено, що статистичними параметрами стилів є частоти голосних, приголосних, пропуски між словами, а також м'яких і сонорних груп приголосних.

2. Аналіз літературних даних і постановка проблеми

Для автоматичного визначення мови аналізують відформатовані уривки тексту: літери розташовані за спаданням частот їх появи в уривку (частоти подаються), розрізнення на малу та велику літери не здійснюють. Проаналізувати дані та визначити авторську мову відформатованих уривків можна трьома методами через дослідження [9]:

1) частот голосних і приголосних у тексті;

2) сонорних, дзвінких і глухих приголосних та їх оцінок;

3) частотності вживання літер мови.

Для дослідження особливостей авторського стилю визначають та аналізують коефіцієнти лексичного авторського мовлення. До них відносять зв'язність мовлення, лексичну різноманітність, синтаксичну складність, індекси концентрації та винятковості для авторського уривку та іншого аналізованого уривку. Далі досліджують внутрішню «динаміку» тексту через аналіз цих коефіцієнтів та визначають степень належності аналізованого тексту конкретному авторові [10].

Для визначення степеня плагіату складають звідну групову таблицю. Туди вносять обчислені групові середні значення зв'язності мовлення, лексичної різноманітності та синтаксичної складності, а також індексів концентрації та винятковості для множин подібних за замісто текстів [11]. Обчислюють зону стандартних відхилень і оцінюють таким чином лексичну подібність кожного аналізованого тексту у порівнянні із еталонним [12].

Визначення автора тексту або ідентифікації тексту провадять за результатами аналізу його відформатованого уривку [13]. Слововживання розташовані за спаданням частот їх появи в уривку. Вказано вид мови, до якої належить слововживання (авторська чи не авторська мова). Власні назви видалені з тексту уривку. Спираючись на частотні словники, визначають автора уривку або й сам твір, якщо це можливо [14]. Аналіз феномену авторства полягає у визначенні відмінності між стилями письменників [15]. Це робить мову автора динамічною, захопливою, легкою до сприйняття, які характеристики є індивідуальними, а що можна вважати спільним [16]. Визначають та аналізують степень авторської атрибуції: достовірності, автентичності художнього твору, його автора, місця й часу створення на підставі аналізу стилістичних і технологічних особливостей [17].

Також аналізують динаміку зміни авторського стилю. Із літературних надбань авторів творів, написаних однією мовою та одного часового періоду, вибираються пари тематичних творів, кожну наступну пару вибирають із кроком у h років [6]. Для кожного набору творів необхідно опрацювати по 1000 слововживань із кожного та знайти, скільки з цих слів належать до 100-слівного списку Сводеша. Це інструмент для оцінки ступеню спорідненості між різними мовами/мовленнями за такою ознакою, як подібність найсталішого базового словника; це перелік базових лексем певної мови/мовленя, що його відсортовано за зменшенням їхньої «базовості». Мінімальний набір найважливішої («ядерної») лексики міститься в списку Сводеша зі 100 слів. Використовують також списки з 200 та 207 слів. Порівняння результатів, отриманих у межах групи, дає змогу виявити тенденцію до збільшення (зменшення) кількості спільних слів зі списку Сводеша в роботах цих авторів. Також це визначає їх розбіжність для визначення авторства у спільних публістичних/наукових роботах [6].

Проблема встановлення авторства анонімних та псевдонімних текстів пов'язана як з історико-філологічними, так і з природничо-технічними науками, серед яких особливої ваги у вирішенні питання набирає статистика та теорія ймовірностей. Причому постановка задачі та використання результатів стосуються літературознавчої сфери, а апарат та методи отримання результату – математичної сфери, що вимагає застосування сучасних наукових теорій та обчислювальних засобів [18]. Для опису індивідуального стилю застосовують лінгвоматематичні методи, що сприяє накопиченню даних про властивості одиниць мови та формуванню спеціального наукового апарату атрибуції текстів. За його допомогою стилеметрія бере

участь у розв'язанні 4-ох основних груп практичних задач [3, 19].

1. *Дослідження публікацій або історичних фактів.* Варто лише згадати «Шекспірівське питання», з приводу якого науковці світу сперечаються і досі, починаючи з 1785 р, відколи преподобний Джеймс Уілмот висловив припущення про те, що справжнім автором п'єс Шекспіра був Френсіс Бекон. Також дослідники твердять, що Мольєру належать не всі приписувані йому твори, дискусійним є авторство «Тихого Дону», існує цілий ряд анонімних творів з невідомим або спірним авторством – методики авторської атрибуції допомагають вирішити ці питання. З історичного погляду необхідно пов'язати різні архівні документи з автором і періодом їх написання. Лише в цьому випадку можна робити висновки на основі вмісту історичних текстів.

2. *Сфера освіти, науки і психології.* Із розвитком Інтернет дослідники все менше працюють самостійно, використовуючи вже готові роботи або їх уривки. Цитовані уривки тексту нерідко перевищують вклад автора і часто не містять вказівки на першоджерело. Методами визначення авторства можна виявити подібний плагіат, тим самим здійснити контроль та оцінити належним чином цю роботу [11, 20]. Аналогічно це стосується і наукових робіт, не лише у визначенні коефіцієнтів унікальності тексту (копірайту та рерайту), а також відсотків автоського вкладу у спільних роботах акторського колективу.

3. *Судова практика.* Об'єктами досліджень є питання авторського права та плагіату, письмові свідчення очевидців чи свідчення, зроблені під тиском, а також договори, заповіти, анонімні листи тощо. Одним із найсучасніших напрямів авторської атрибуції є визначення творців комп'ютерних вірусів. Перспективою у розвідках авторської атрибуції текстів є, наприклад, дослідження збереження авторського стилю у перекладах текстів [3, 21].

4. *Кібербезпека.* Визначення стилю автора із швидким розвитком ІТ та активності користувачів Інтернет досить важливе при ідентифікації шахраїв через їх історію в соціальних мережах. Це не лише допомогає знаходити провопорушників, але може сприяти запобіганню злочинів (наприклад, активність зловісної організації «Синій кіт» або діяльність в мережах так званих тролів у відомій інформаційній війні між словянськими державами).

3. Мета та задачі дослідження

Метою роботи є розроблення формального підходу визначення стилю автора у україномовних текстах на основі технології статистичної лінгвістики.

Для досягнення мети сформульовані такі завдання:

– розробити метод визначення стилю автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту;

– розробити формальний підхід до проектування програмного забезпечення контент-моніторінгу для визначення стилю автора в україномовних текстах на основі Web Mining та лексичного аналізу визначенних стопових слів текстового контенту;

– отримати та проаналізувати результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення стилю автора в україномовних наукових текстах технічного профілю.

4. Метод визначення стилю автора текстового контенту

Лінгвостатистичне підгрунтя для здійснення дослідження з метою атрибуції тексту складають [3, 18–24]:

1) первинне опрацювання лінгвістичних даних (побудова рядів розподілу, обчислення статистик, статистичних оцінок та інші параметри лінгвометрії);

2) лексикографічне опрацювання текстових даних (створення частотних і алфавітно-частотних словників, словників-конкордансів, слововказівників, зворотних словників, словників ключових слів стилю письменника тощо.

Застосування методик лінгвометрії для статистичного опису тексту дає змогу виконувати дослідження, що стосуються феномену авторства [25]. Метод аналізу та інтерпретації на лінгвістичному рівні стилістичних особливостей і закономірностей письменницького стилю певного автора (або певної літературної епохи) використовує алгоритм 1.

Алгоритм 1. Аналіз та інтерпретація на лінгвістичному рівні стилістичних особливостей і закономірностей письменницького стилю певного автора.

Етап 1. *Відбір текстів.* Важливим є спосіб організації відбору та обсяг текстової вибірки: для визначення характеристик він повинен складати щонайменше 18 тис. слів [23–25].

Етап 2. *Лематизація текстових одиниць.* Об'єднання словоформ під лемою мови [5].

Етап 3. *Усунення неоднорідності текстових одиниць.* Розв'язання проблеми неоднорідності текстових одиниць, наприклад, із погляду їх відношення до різних видів мови (авторська, не авторська і т. п.).

Етап 4. *Побудова системи, організація на їх основі статистичних розподілів у потрібних частотних словників шкалах.* Частотний словник – тип словника, де наведено кількість вживань (частоту) певної одиниці мови (складу, слова, словоформи, словосполучення, ідіоми, фразеологізму) в різних текстах певного обсягу. Зазвичай, подають абсолютну та відносну частоту вживання мовних одиниць, словникові статті розміщують за спаданням частот [3].

Етап 5. *Пошук параметрів, що адекватно відображають структуру частотного словника.* Кількість параметрів є різноманітною, наприклад, для опису французьких текстів XVII ст. запропоновано 51 параметр [25]. Знайдені загальні параметри в [26–31] дають змогу сформулювати кілька основних лінгвостатистичних методів дослідження тексту:

– метод опорних слів (підрахунок загальної частоти вживання та знаходження відсоткового складу службових слів [18–22]: прийменників, сполучників, часток);

– метод розділових знаків (підрахунок лише кількості внутрішніх і зовнішніх розділових знаків);

– метод слів (підрахунок лише слів певної довжини);

– метод речень (підрахунок лише речень визначеної довжини);

– синтаксичний метод (підрахунок розділових знаків, слів і речень певної довжини);

– комбінований (поєднання методів опорних слів і синтаксичного).

Етап 6. *Перевірка параметрів на ефективність.* Застосування загальних методів перевірки відібраних параметрів на ефективність.

Етап 7. *Математичне моделювання лексикостатистичних розподілів.* Застосування загальних методів математичного апарату моделювання лексикостатистичних розподілів.

Етап 8. *Побудова статистичних класифікацій* (авторських еталонів), що відображають стилістичні закономірності в межах творів певного автора чи певної літературної епохи (або послідовності літературних епох).

Етап 9. *інтерпретація отриманих результатів* із позицій історико-літературних уявлень, загальної й історичної стилістики.

З використанням алг. 1 вирішують завдання авторської атрибуції, яке можна сформулювати, наприклад, наступним чином. Нехай існує статистично опрацьований доробок автора (еталон). Необхідно оцінити належність певних уривків до еталону із застосуванням відповідних методів. Розглянемо для ілюстрації творчість Автора 1 та його публікації [24]. Причому будемо вважати, що авторський еталон вже побудований – завдання з відбору текстів, лематизації та проблеми неоднорідності вирішені, опрацьований матеріал сформований у вигляді частотного словника [3]. Використаємо для атрибуції метод опорних слів, результати подамо у вигляді коефіцієнтів кореляції та графічно. Окремо згадаємо про еволюцію значущості одного із параметрів тексту – службових слів – в авторській атрибуції текстів (табл. 1).

Для індивідуального стилю письменника показовими є саме службові слова, оскільки вони ніяк не пов'язані з темою і змістом книги [3]. Вважатимемо вказаний параметр дослідження тексту ефективним та приймемо список стоп-слів (службових слів) [25], викладений у табл. 1 (усього 71 слово).

Таблиця 1

Службові частини української мови (стоп-слова)

Частина мови	Список стопових слів
Прийменники	в, на, з, за, до, по, у, біля, від, для, без, про, через, при, над, з-за, з-під, під, близько, вглиб, крізь, поза, проміж
Сполучники	і, й, що, так, хоча, коли, або, щоб, якщо, також, чи, тобто, проте, немов, а, але, та, через те що, одначе, та й
Частки	не, так, ж, же, навіть, би, або, лише, то, ні, адже, он, тобто, уже, чи, аякже, це, тільки, ось, ледве чи, мов, немов

5. Результати досліджень визначення стилю автора в україномовних текстах на основі технології статистичної лінгвістики

Проаналізовано 100 наукових публікацій з двох номерів (783 та 805) Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі». Розглянемо довільні чотири уривки з проаналізованих текстів, відформатовані з огляду на вибір методу атрибуції: з кожного уривку вибрано лише прийменники, сполучники та частки. Наведено загальну кількість слововживань в уривку, власні назви не враховуються. У табл. 2 для кожного з уривків вказано абсолютну частоту (АЧ) та відносну частоту (ВЧ) появи службового слова, а також відносну частоту появи вказаного слова в еталоні.

Таблиця 2

Абсолютні та відносні частоти появи стопових слів в Уривку та еталоні

Уривок	Стоп-слово	АЧ	ВЧ	Частина мови	ВЧ в еталоні
1	2	3	4	5	6
1 (107 слів)	але	1	0,0093	Сполучник	0,0074
	в	2	0,0187	Прийменник	0,0140
	для	3	0,0280	Прийменник	0,0024
	до	1	0,0093	Прийменник	0,0113
	з	1	0,0093	Прийменник	0,0129
	і	14	0,1308	Сполучник	0,0300
	й	1	0,0093	Сполучник	0,0038
	мов	1	0,0093	Частка	0,0022
	не	2	0,0187	Частка	0,0237
	про	2	0,0187	Прийменник	0,0040
	та	2	0,0187	Сполучник	0,0047
	що	1	0,0093	Сполучник	0,0206
2 (117 слів)	а	2	0,0171	Сполучник	0,0116
	в	3	0,0256	Прийменник	0,0140
	від	1	0,0085	Прийменник	0,0034
	до	1	0,0085	Прийменник	0,0113
	ж	1	0,0085	Сполучник	0,0033
	з	2	0,0171	Прийменник	0,0129
	за	1	0,0085	Прийменник	0,0053
	і	2	0,0171	Сполучник	0,0300
	й	2	0,0171	Сполучник	0,0038
	на	1	0,0085	Прийменник	0,0159

Продовження таблиці 2

1	2	3	4	5	6
2 (117 слів)	над	1	0,0085	Прийменник	0,0005
	не	2	0,0171	Частка	0,0237
	ні	1	0,0085	Частка	0,0024
	ось	1	0,0085	Частка	0,0012
	от	1	0,0085	Частка	0,0005
	се	1	0,0085	Частка	0,0074
	хіба	1	0,0085	Частка	0,0006
	хоч	1	0,0085	Частка	0,0010
	що	2	0,0171	Сполучник	0,0206
	як	1	0,0085	Сполучник	0,0060
3 (162 слів)	а	4	0,0247	Сполучник	0,0116
	але	2	0,0123	Сполучник	0,0074
	без	1	0,0062	Прийменник	0,0008
	бо	1	0,0062	Сполучник	0,0012
	в	1	0,0062	Прийменник	0,0140
	від	1	0,0062	Прийменник	0,0034
	ж	1	0,0062	Сполучник	0,0033
	з	4	0,0247	Прийменник	0,0129
	за	2	0,0123	Прийменник	0,0053
	і	1	0,0062	Сполучник	0,0300
	й	4	0,0247	Сполучник	0,0038
	на	6	0,0370	Сполучник	0,0159
	навіть	2	0,0123	Частка	0,0011
	не	3	0,0185	Частка	0,0237
	під	4	0,0247	Прийменник	0,0011
	таки	1	0,0062	Частка	0,0004
	тож	1	0,0062	Сполучник	0,0001
	у	4	0,0247	Прийменник	0,0088
	що	3	0,0185	Сполучник	0,0206
	щоб	1	0,0062	Сполучник	0,0028
	як	1	0,0062	Сполучник	0,0060
4 (149 слів)	адже	1	0,00671	Частка	0,0011
	але	2	0,01342	Сполучник	0,0074
	би	1	0,00671	Частка	0,0033
	в	1	0,00671	Прийменник	0,0140
	ж	1	0,00671	Сполучник	0,0033
	з	3	0,02013	Прийменник	0,0129
	за	1	0,00671	Прийменник	0,0053
	і	4	0,02685	Прийменник	0,0300
	мов	1	0,00671	Частка	0,0022
	на	7	0,04698	Прийменник	0,0159
	не	4	0,02685	Частка	0,0237
	отсе	1	0,00671	Частка	0,0003
	при	1	0,00671	Прийменник	0,0018
	про	2	0,01342	Прийменник	0,0040
	се	1	0,00671	Частка	0,0074
	у	2	0,01342	Прийменник	0,0088
	чи	2	0,01342	Сполучник	0,0027
	що	7	0,04698	Сполучник	0,0206
	щоб	1	0,00671	Сполучник	0,0028
	як	1	0,00671	Сполучник	0,0060

На рис. 1 подане графічне зображення відносної частоти появи стопових слів в Уривку 1 та в еталоні. Коефіцієнт кореляції для службових слів у цьому випадку складає $R_{e\text{-}y1}$=0,6076. Графічне зображення відносної частоти появи службових слів в Уривку 2 та в еталоні подане на рис. 2. Коефіцієнт кореляції для службових слів у цьому випадку складає $R_{e\text{-}y2}$=0,7066.

Графічне зображення відносної частоти появи службових слів в Уривку 3 та в еталоні подане на

рис. 3. Коефіцієнт кореляції для службових слів у даному випадку складає $R_{e\text{-}y3}$=0,2810.

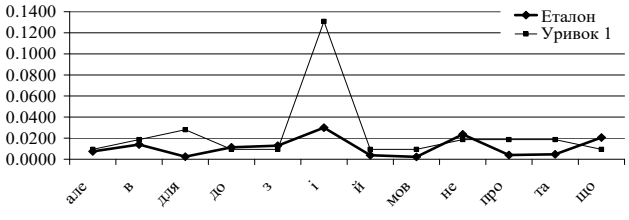

Рис. 1. Відносна частота появи службових слів в Уривку 1 та в еталоні

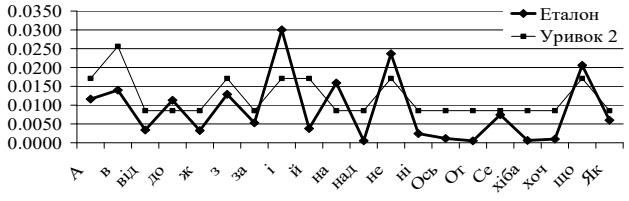

Рис. 2. Відносна частота появи службових слів в Уривку 2 та в еталоні

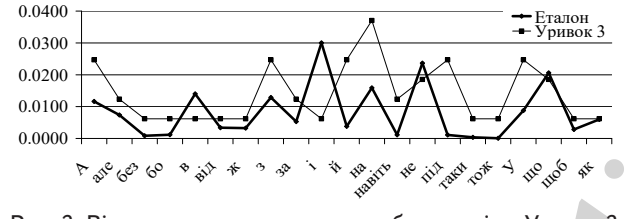

Рис. 3. Відносна частота появи службових слів в Уривку 3 та в еталоні

Графічне зображення відносної частоти появи службових слів в Уривку 4 та в еталоні подане на рис. 4. Коефіцієнт кореляції для службових слів у цьому випадку складає $R_{e\text{-}y4}$=0,7326.

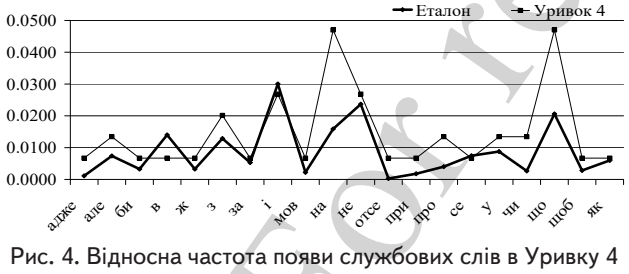

Рис. 4. Відносна частота появи службових слів в Уривку 4 та в еталоні

Наведемо також коефіцієнти кореляції для кожного зі службових слів для уривків 1–4 (табл. 3).

Таблиця 3

Коефіцієнти кореляції для службової частини мови

Уривок	Прийменник	Сполучник	Частка
1	$R_{e\text{-}y1Z}$=0,72	$R_{e\text{-}y1S}$=0,79	$R_{e\text{-}y1C}$=1
2	$R_{e\text{-}y2Z}$=0,4928	$R_{e\text{-}y2S}$=0,5714	$R_{e\text{-}y2C}$=0,9580
3	$R_{e\text{-}y3Z}$=0,1517	$R_{e\text{-}y3S}$=0,1624	$R_{e\text{-}y3C}$=0,8800
4	$R_{e\text{-}y4Z}$=0,5639	$R_{e\text{-}y4S}$=0,9544	$R_{e\text{-}y4C}$=0,9594

Аналізуючи коефіцієнти кореляції для службових слів, приходимо до висновку, що ймовірність належності уривків до досліджуваного еталону найбільшою є для Уривку 4, за ним – Уривок 2, Уривок 1, Уривок 3.

Зауважимо, що для всіх чотирьох уривків простежуються стабільно високі коефіцієнти кореляції для часток, що можемо розуміти як відсутність впливу часток на авторський стиль. Додатково для уривків проаналізуємо частотності поява лише прийменників і сполучників, знайдемо відповідні коефіцієнти кореляції та порівняємо результати (табл. 4).

Таблиця 4

Коефіцієнти кореляції для кожного з уривків

Уривок	Уривок 1	Уривок 2	Уривок 3	Уривок 4
Коефіцієнт $R_{e\text{-}y}$	$R_{e\text{-}y1}$= =0,6076	$R_{e\text{-}y2}$= =0,7066	$R_{e\text{-}y3}$= =0,2810	$R_{e\text{-}y4}$= =0,7326
Коефіцієнт $R'_{e\text{-}y}$	$R'_{e\text{-}y1}$= =0,6900	$R'_{e\text{-}y2}$= =0,4913	$R'_{e\text{-}y3}$= =0,2254	$R'_{e\text{-}y4}$= =0,6905

Уривок 4 так і залишився найімовірнішим кандидатом на належність його до еталону, а наступним із незначним відривом став Уривок 1, далі – Уривок 2. Уривок 3, як і у попередньому дослідженні, має найменшу ймовірність належати до еталону. Для підтвердження результатів звернемося до аналізованих текстів, з яких узято уривки для дослідження.

Отже, застосування методу опорних слів дало такі результати: серед досліджуваних уривків найбільшу ймовірність належати до еталону справді отримав той уривок, що належить до аналізованих текстів. Інші результати також підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Так, у першому дослідженні наступну за величиною ймовірність належати до еталону має уривок з іншого твору того самого автора. Уривок 1, що теж належить до еталону, «програв» Уривку 4 лише одну десяту в коефіцієнті кореляції. Також адекватним є результат для Уривку 3, якого з еталоном розділяють близько ста років. Висунуте припущення в [25] про незначущість впливу частки як параметра методу на результати привело до зменшення коефіцієнтів кореляції, але розташувало ймовірність належати до еталону для уривків у вірному порядку. Понад усе, різниця між коефіцієнтами кореляції для Уривку 1 та Уривку 4 значно зменшилася і склала 0,0005. Проте, для підтвердження чи спростування того факту, що частки не є визначальним фактором в авторському стилі необхідно виконати грунтовніші дослідження.

6. Обговорення результатів дослідження аналізованих україномовного контенту для визначення стилю автора

Для досягнення мети дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі Victana [24]. Аналіз статистики функціонування системи виявлення множини стопових слів із 100 наукових статей технічного спрямування проведено у 3 етапи (алг. 2).

Алгоритм 2. Аналіз та інтерпретація лінгвостатистичних досліджень визначення та аналізу стилю автора.

Eman I. Лексичний аналіз тексту для визначення стопових слів та розрахунку коефіцієнтів лексичного авторського мовлення (різноманітності тексту).

Eman II. Визначення стилю автора за методами стилеметрії.

Eman III. Аналіз уривків тексту методами глотохронології, використовуючи списки Сводеша.

Emanu I, II розглянуті в попередньому розділі статті. Розглянемо етап III.

Основним завданням є визначення кількості слів із 200-слівного списку Сводеша, які присутні в творах різних часових зрізів, та визначення відсоткового складу таких слів в уривках. Також дослідимо кількість спільних слів зі списку Сводеша для обраних уривків. Для розгляду підберемо фрагменти, написані з розривом у кілька років. Нехай уривки складатимуться, наприклад, із 250 слів, не враховуючи заголовка та власних назв. Порівняння 200-слівного списку Сводеша та Уривку 1 з аналізованих текстів викладені у табл. 5.

Таблиця 5

Слова зі списку Сводеша в Уривку 1

№	Слово	Абослютна частота	Відносна частота
1	все	4	0,0526
2	і	19	0,2500
3	на	3	0,0395
4	он	5	0,0658
5	слухати	1	0,0132
6	як	2	0,0263
7	я	6	0,0789
8	в	4	0,0526
9	знати	2	0,0263
10	довго	2	0,0263
11	чоловік	1	0,0132
12	багато	1	0,0132
13	ім'я	1	0,0132
14	ні	3	0,0395
15	старий	2	0,0263
16	сонце	1	0,0132
17	що	6	0,0789
18	там	3	0,0395
19	what	1	0,0132
20	який	2	0,0263
21	з	5	0,0658
22	рік	1	0,0132
23	ви	1	0,0132
Усього		76	

В Уривку 1, обсягом 253 слова, є 23 слова з 200-слівного списку Сводеша. Ці слова складають 30,04 % від усього уривку. Уривок 2 – це фрагмент з аналізованих текстів. Порівняння 200-слівного списку Сводеша та Уривку 2 викладені у табл. 6.

В Уривку 2, обсягом 262 слова, є 24 слова з 200-слівного списку Сводеша. Ці слова складають 18,7 % від усього уривку. Уривок 3 – це фрагмент з аналізованих текстів. Порівняння 200-слівного списку Сводеша та Уривку 3 викладені у табл. 7.

Таблиця 6

Слова зі списку Сводеша в Уривку 2

№	Слово	Абослютна частота	Відносна частота
1	все	4	0,0816
2	і	6	0,1224
3	на	1	0,0204
4	назад	1	0,0204
5	далеко	1	0,0204
6	товстий	1	0,0204
7	потік	1	0,0204
8	тут	2	0,0408
9	якщо	1	0,0204
10	в	7	0,1429
11	знати	2	0,0408
12	ні	1	0,0204
13	один	2	0,0408
14	інший	1	0,0204
15	дещо	1	0,0204
16	що	3	0,0612
17	там	2	0,0408
18	це	2	0,0408
19	кидати	1	0,0204
20	який	4	0,0816
21	білий	1	0,0204
22	хто	1	0,0204
23	з	2	0,0408
24	ви	1	0,0204
Усього слів		49	

Таблиця 7

Слова зі списку Сводеша в Уривку 3

№	Слово	Абослютна частота	Відносна частота
1	все	3	0,0652
2	і	10	0,2174
3	на	1	0,0217
4	прихо-дити	1	0,0217
5	тут	1	0,0217
6	якщо	1	0,0217
7	в	4	0,087
8	знати	2	0,0435
9	довго	1	0,0217
10	ні	7	0,1522
11	інший	1	0,0217
12	казати	1	0,0217
13	що	4	0,087
14	там	1	0,0217
15	вони	2	0,0435
16	це	1	0,0217
17	який	1	0,0217
18	хто	2	0,0435
19	з	2	0,0435
Усього слів		46	

В Уривку 3, обсягом 246 слів, є 19 слів із 200-слівного списку Сводеша. Ці слова складають 18,7 % від

усього уривку. Аналізуючи отримані дані, зауважуємо, що слова зі списку Сводеша в Уривку 1 складають 30 % від уривку, що значно більше, ніж 18,7 %, як в Уривках 2 та 3 (рис. 5). Такі результати є закономірними та прозорими: з часом збагачується і словниковий запас людини. Також для цих уривків на рис. 6 графічно зображено такі результати:

– у вузлах зазначено уривок та кількість слів у ньому зі списку Сводеша;

– на дугах вказано кількість спільних слів зі списку Сводеша для цих уривків та коефіцієнт кореляції для цих уривків;

– у центрі зазначена загальна кількість слів, спільних для уривків та списку Сводеша (табл. 8).

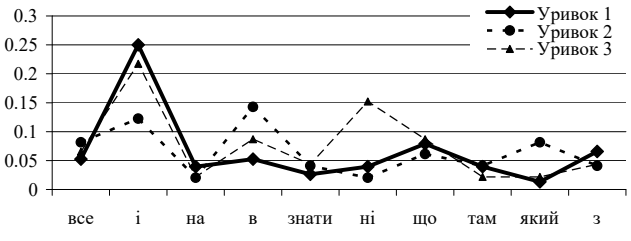

Рис. 5. Чисельні результати дослідження уривків

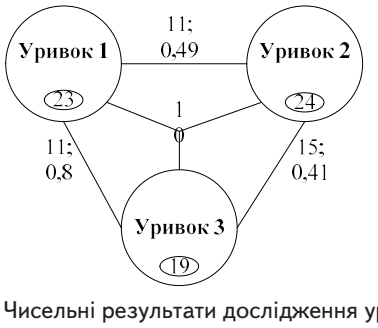

Рис. 6. Чисельні результати дослідження уривків

Таблиця 8

Слова, спільні для Уривків 1–3 та списку Сводеша

№	Спільні слова	Відносна частота в Уривку 1	Відносна частота в Уривку 2	Відносна частота в Уривку 3
1	все	0,0526	0,0816	0,0652
2	і	0,25	0,1224	0,2174
3	на	0,0395	0,0204	0,0217
4	в	0,0526	0,1429	0,087
5	знати	0,0263	0,0408	0,0435
6	ні	0,0395	0,0204	0,1522
7	що	0,0789	0,0612	0,087
8	там	0,0395	0,0408	0,0217
9	який	0,0132	0,0816	0,0217
10	з	0,0658	0,0408	0,0435

Обсяг проведених досліджень не дає змоги стверджувати, що такий високий коефіцієнт кореляції, як між Уривком 1 та Уривком 3, є закономірним. Коефіцієнт наразі дозволяє висунути гіпотезу про те, що, загалом Уривок 1 або написаний в інший часовий проміжок, ніж Уривки 2, 3, або написані іншою особою. Той факт, що така залежність справді існує, чи що це випадковий збіг через невдало обраний уривок, потребує значно ширших досліджень.

7. Висновки

1. Розроблено метод визначення стилю автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Метод полягає в порівняльному аналізі авторської атрибуції в статистично опрацьованому доробку автора (еталоні) з довільним аналізованим уривком. Метод оцінює належність певних уривків до еталону із аналізом відповідних коефіцієнтів лексичного авторського мовлення. Причому метод працює при умові, що авторський еталон вже побудований та проаналізований – завдання з відбору текстів, лематизації та проблеми неоднорідності вирішені, опрацьований матеріал сформований у вигляді частотного словника службових слів (стопових слів). Для атрибуції використано метод опорних слів, результати подано у вигляді коефіцієнтів кореляції. Окремо згадаємо про еволюцію значущості одного із параметрів тексту – в авторській атрибуції текстів.

Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Для індивідуального стилю письменника показовими є саме службові слова, оскільки вони ніяк не пов'язані з темою і змістом книги. Аналізовані уривки, відформатовані з огляду на вибір методу атрибуції: з кожного уривку автоматично обрано лише прийменники, сполучники та частки. Підраховано загальну кількість слововживань в уривку, власні назви не враховуються. Для кожного з уривків проаналізовані та порівняні із еталонним значеннями абсолютні та відносні частоти появи стопових слова. Отже, застосування методу опорних слів дає такі результати: знаходження серед досліджуваних уривків того, що найбільш ймовірно належить до еталону. Інші результати також підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Висунуте припущення про незначущість впливу частки як параметра методу на результати привело до зменшення коефіцієнтів кореляції, але розташувало ймовірність належати до еталону для уривків у вірному порядку. Проте, для підтвердження чи спростування того факту, що частки не є визначальним фактором в авторському стилі необхідно виконати грунтовніші дослідження.

Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Особливостями алгоритму є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Наведено теоретичне та експериментальне обгрунтування методу контент-моніторінгу та визначення стопових слів україномовного тексту. Метод спрямовано на автоматичне виявлення значущих стопових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації парсингу контенту.

2. Запропоновано підхід до розроблення програмного забезпечення контент-моніторінгу для визначення стилю автора в україномовних текстах на основі Web Mining. Особливість підходу полягає у адаптації лінгвостатистичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів.

3. Досліджено результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення стилю автора в україномовних наукових текстах технічного профілю. Досліджено 100 наукових публікацій з двох номерів (783 та 805) Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі». Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення стилю автора з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

Література

1. Анисимов, А. Система обработки текстов на естественном языке [Текст] / А. Анисимов, А. Марченко // Искусственный интеллект. – 2002. – № 4. – С. 157–163.
2. Перебийніс, В. Математична лінгвістика. Українська мова [Текст] / В. Перебийніс. – К.: Українська енциклопедія, 2000. – С. 287–302.
3. Бук, С. Н. Основи статистичної лінгвістики [Текст] / С. Н. Бук; ред. Ф. С. Бацевич. – Л.: Видавничий центр ЛНУ ім. І. Франка, 2008. – 124 с.
4. Варфоломеев, А. П. Психосемантика слова и лингвостатистика текста [Текст] / А. П. Варфоломеев. – Калининград: КГУ, 2000. – 37 с.
5. Когнитивная стилометрия: к постановке проблемы [Электронный ресурс]. – Режим доступа: http://www.manekin.narod.ru/hist/styl.htm
6. Дьячок, М. Т. Глоттохронология: пятьдесят лет спустя [Текст] / М. Т. Дьячок // Сибирский лингвистический семинар. – 2002. – № 1. – С. 44–69.
7. Перебийніс, В. І. Статистичні методи для лінгвістів [Текст] / В. І. Перебийніс. – Вінниця: Нова книга, 2013. – 176 с.
8. Кочерган, М. П. Вступ до мовознавства [Текст] / М. П. Кочерган. – К.: Академія, 2005. – 329 с.
9. Сушко, С. Частоти повторюваності букв і біграм у відкритих текстах українською мовою [Текст] / С. Сушко, Л. Фомичова, Є. Барсуков // Захист інформації. – 2010. – Т. 12, № 3. doi: 10.18372/2410-7840.12.1968
10. Хмелев, Д. Как определить писателя? [Электронный ресурс] / Д. Хмелев // Компьютерра-Онлайн. – 2000. – Режим доступа: http://old.computerra.ru/2000/338/195699/
11. Ланде, Д. В. Підхід до рішення проблем пошуку двомовного плагіату [Текст] / Д. В. Ланде, В. В. Жигало // Проблеми інформатизації та управління. – 2008. – № 2 (24). – С. 125–129.
12. Морозов, Н. А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд [Электронный ресурс] / Н. А. Морозов // Известия отд. русского языка и словесности Имп. Акад.наук. – 1915. – Т. XX. – Режим доступа: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
13. Бублейник, Л. В. Особливості художнього мовлення [Текст] / Л. В. Бублейник. – Луцьк: Вежа, 2000. – 179 с.
14. Родионова, Е. С. Методы атрибуции художественных текстов [Текст] / Е. С. Родионова // Структурная и прикладная лингвистика. – 2008. – Вып. 7. – С. 118–127.
15. Мещеряков, Р. В. Модели определения авторства текста [Текст] / Р. В. Мещеряков, Н. С. Васюков // Измерения, автоматизация и моделирование в промышленности и научных исследованиях. – 2005. – С. 25–29. – Режим доступа: http://db.biysk.secna.ru/conference/conference.conference.doc_download?id_thesis_dl=427
16. Khomytska, I. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level [Text] / I. Khomytska, V. Teslyuk // Advances in Intelligent Systems and Computing. – 2016. – P. 149–163. doi: 10.1007/978-3-319-45991-2_10
17. Khomytska, I. Specifics of phonostatistical structure of the scientific style in English style system [Text] / I. Khomytska, V. Teslyuk // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589887
18. Lytvyn, V. Classification Methods of Text Documents Using Ontology Based Approach [Text] / V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H. Rishnyak // Advances in Intelligent Systems and Computing. – 2016. – P. 229–240. doi: 10.1007/978-3-319-45991-2_15
19. Lytvyn, V. The method of formation of the status of personality understanding based on the content analysis [Text] / V. Lytvyn, P. Pukach, I. Bobyk, V. Vysotska // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 5, Issue 2 (83). – P. 4–12. doi: 10.15587/1729-4061.2016.77174
20. Vysotska, V. Linguistic analysis of textual commercial content for information resources processing [Text] / V. Vysotska // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). – 2016. doi: 10.1109/tcset.2016.7452160
21. Vysotska, V. Information technology of processing information resources in electronic content commerce systems [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589909
22. Vysotska, V. The commercial content digest formation and distributional process [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589902
23. Марченко, О. О. Моделювання семантичного контексту при аналізі текстів на природній мові [Текст] / О. О. Марченко // Вісник Київського університету. – 2006. – № 3. – С. 230–235.

24. Блог Вікторії Анатоліївни [Електронний ресурс]. – Режим доступу: http://victana.lviv.ua/index.php/kliuchovi-slova
25. Родионова, Е. С. Методы атрибуции художественных текстов [Текст] / Е. С. Родионова // Структурная и прикладная лингвистика. – 2008. – Вып. 7. – С. 118–127. – Режим доступа: http://epir.ru/pragmat!/projects/corneille/files/Metody_atributsii.pdf
26. Lytvyn, V. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining [Text] / V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, D. Ugryn // Eastern-European Journal of Enterprise Technologies. – 2017. – Vol. 2, Issue 2 (86). – P. 14–23. doi: 10.15587/1729-4061.2017.98750
27. Lytvyn, V. Content linguistic analysis methods for textual documents classification [Text] / V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H. Rishnyak // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589903
28. Lytvyn, V. Designing architecture of electronic content commerce system [Text] / V. Lytvyn, V. Vysotska // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). – 2015. doi: 10.1109/stc-csit.2015.7325446
29. Vysotska, V. Analysis features of information resources processing [Text] / V. Vysotska, L. Chyrun // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). – 2015. doi: 10.1109/stc-csit.2015.7325448
30. Chen, J. Smart Data Integration by Goal Driven Ontology Learning [Text] / J. Chen, D. Dosyn, V. Lytvyn, A. Sachenko // Advances in Big Data. Proceedings of the 2nd INNS Conference on Big Data. – October 23-25, 2016. – Thessaloniki, Greece. – P. 283-292.
31. Mykhailiuk, A. A Creation of the Linguistic Ontology Based on a structured Electronic Encyclopedic Resource [Text] / A. Mykhailiuk, O. Mykhailiuk, O. Pylypchuk, V. Tarasenko // International Journal of Computing. – 2012. – Vol. 11, Issue 3. – P. 191-202.