УДК 004.89
DOI: 10.15587/1729-4061.2019.186834

Розроблення квантитативного методу автоматичного визначення авторства текстового контенту на основі статистичного аналізу розподілу N-грам

В. В. Литвин, В. А. Висоцька, І. С. Будз, Я. М. Пелех, Н. Б. Сокульська, Р. А. Ковальчук, Л. В. Дзюбик, О. В. Терещук, М. П. Комар

Розглянуто особливості застосування технологій лінгвостатистики для ідентифікації стилістики автора текстового контенту науково-технічного профілю. Квантитативний лінгвістичний аналіз тексту використовує переваги контент-моніторінгу на основі методів NLP для визначення та аналізу множини стопових слів, ключових слів, стійких словосполучень та дослідження N-грам. Останні використовують в методах лінгвометрії для визначення приналежності аналізованого тексту конкретному авторові у відсотках. Розроблено квантитативний метод автоматичного визначення авторства текстового контенту на основі статистичного аналізу розподілу 3-грам. Запропоновано підхід реалізації визначення автора україномовного тексту науково-технічного профілю. Отримано експериментальні результати запропонованого методу для визначення приналежності аналізованого тексту конкретному автору за наявності еталонного авторського тексту. Застосування лінгвостатистичного аналізу 3-грам до множини статей дозволить сформувати підмножину подібних за лінгвістичними характеристиками публікацій. Накладання на підмножину додаткових умов у вигляді проведення статистичних та квантитативних аналізів (множини ключових слів, стійких словосполучень, стилеметричного, лігвометричного тощо) дозволить значно скоротити цю підмножину, уточнивши список ймовірніших авторських робіт. Для якісного та ефективного аналізу контенту при визначенні ступеня авторства конкретному автору пропонуємо аналізувати еталонного тексту та досліджуваного в декілька етапів: лінгвометричний аналіз коефіцієнтів різноманіття авторського мовлення, стилометричний аналіз, аналіз стійких словосполучень, лінгвостатистичний аналіз 3-грам. Для автоматизованого опрацювання тексту має велике значення не тільки частота появи тієї чи іншої категорії, а взагалі присутність в досліджуваному тексті. Кількісний підрахунок дозволяє зробити об'єктивні висновки щодо спрямованості матеріалів за кількістю уживань одиниць аналізу в досліджуваних текстах. Якісний аналіз робить те саме, але внаслідок дослідження того, чи зустрічається (і в якому контексті) певна важлива оригінальна категорія взагалі.

Ключові слова: NLP, контент, контент-моніторінг, стоп-слова, контент-аналіз, статистийний лінгвістичний аналіз, квантитативна лінгвістика, статистична лінгвістика, лінгвометрія

1. Вступ

У зв'язку зі збільшенням доступності та поширення текстового контенту в Інтернет зростає ступінь важливості використання автоматичних методів аналі-

зу змісту тексту [1]. До завдань аналізу контенту відносять задачі класифікації і кластеризації текстових публікацій за різними критеріями, наприклад, жанру, епосі написання, формату (роман, есе, наукова стаття), емоційному забарвленню, стилю мовлення, а також задача визначення авторства тексту [2]. Із спрощенням доступу до різних даних, розширенням можливості пошуку, копіювання та розповсюдження даних в Інтернет стає актуальною задача ідентифікації автора [3]. Задачі, пов'язані з встановленням авторства, є важливими в лінгвістичних, історичних і криміналістичних дослідженнях [4]. Загальнодоступність електронних пристроїв дозволяє відсунути розпізнавання автора з залученням великої кількості експертів на другий план, прискорити і спростити цей процес за допомогою його автоматизації [5]. Поняття ідентифікація автора визначається як процес встановлення автора за множиною загальних і приватних ознак тексту, що складають авторський стиль [6].

Сьогодні користуються популярністю статистичні методи визначення авторства, засновані на пошуку авторського інваріанту [7]. Це характеристика мовної особливості тексту (лексичної, граматичної, фразеологічної тощо) [8]. Зокрема, інваріантом є частка голосних/приголосних, частота вживання певної частини мови, ймовірність переходів від однієї частини мови до іншої, слова-паразити, інформаційна ентропія тощо [9]. Ефективним є статистичний метод визначення автора і жанру тексту, заснований на розподілі частот літеросполучень (N-грам) [10]. Але точність статистичних методів визначення авторства сильно залежить від специфіки використовуваних даних: від мови стилю мовлення та довжини тексту [11]. В силу цього важко робити висновок про точність такого підходу на науково-технічних статтях [12]. З цієї причини необхідно аналізувати застосовність такого математичного апарату, як розподіл частот літеросполучень для різних мов одночасно з іншими техніками при вирішенні завдання встановлення авторства текстів, що мають різні довжини і написані в різному стилі мовлення [13]. У роботах [1–7] запропоновано та досліджено методи визначення автора україномовного текстового контенту науково-технічного спрямування. Для реалізації цих методів можна використати різні алгоритми [14], зокрема квантитативні [15]. Тому виникає задача аналізу таких алгоритмів з метою пошуку найефективнішого [16]. Авторифікація авторства є технікою визначення автора тексту, коли неоднозначно, хто її написав [17]. Це корисно, коли декілька людей претендують на авторство однієї публікації [18] або у випадках, коли ніхто не претендує на авторство текстового контенту [19], наприклад, так звані тролі в соціальних мережах під час інформаційної війни [20]. Складність проблеми авторського тексту, очевидно, експоненціально вища, більша кількість вірогідних авторів [21]. Наявність авторських текстових зразків також є суттєвою при просуненні цієї проблеми [22]. Атрибуція авторського тексту включає наступні три проблеми [23]:

– виявлення автора текстового автора з групи імовірних або очікуваних авторів, де автор завжди знаходиться у групі підозрюваних [24];

– не ідентифікація автора текстового автора з групи вірогідних або очікуваних авторів, де автор може не бути в групі підозрюваних [25];

– оцінка можливості даного тексту, написаного даним автором чи ні [26].

Тому задача автоматичного визначення автора текстового контенту науково-технічного спрямування є актуальною й потребує нових (досконаліших) підходів до її розв'язування [27].

2. Аналіз літературних даних і постановка проблеми

В роботах [1–3, 28] наведені результати досліджень мовного та мовленнєвого матеріалу на репрезентативному масиві текстів. Це має бути однорідний масив (корпус) певних одиниць, тобто генеральною сукупністю (ГС) [3]. Показано, що обсяг і характер ГС залежать від завдань дослідження. Наприклад, якщо досліджують особливості стилю Івана Франка, то ГС – усі його твори. Якщо досліджують українську мову XX ст., то ГС – усі тексти (мовлені та писані) XX ст. [3] Межі останньої важко виявити точно, а все усне мовлення просто неможливо дослідити [29], особливо при аналізі стилю авторського мовлення. Також залишилися невирішеними питання, пов'язані з визначення авторства тексту в колективних роботах науково-технічного спрямування на основі аналізу еталонів. Причиною є відсутність експериментів у цьому напрямку. Іншою причиною є наявність недостатніх статистичних даних для формування висновків у зв'язку з тим, що автори в цьому напрямку рідко пишуть одноосібні роботи, а деяких напрямках навіть взагалі рідко пишуть колективні роботи. Варіантом подолання відповідних труднощів, коли суцільне обстеження ГС неможливе, роблять вибірку та формують множину параметрів для відповідного аналізу [1-3, 30]. Все це дає підстави стверджувати, що доцільним є проведення дослідження, присвяченого визначення авторства текстового контенту на основі статистичного аналізу розподілу характеристик авторського мовлення пр. достатній вибірці даних.

Вибірка – це певна кількість матеріалу, на підставі дослідження якого можна зробити правильні висновки про всю ГС [31]. Основні вимоги до вибірки: репрезентативність та однорідність [32]. Щоби бути репрезентативною, вибірка повинна [33]:

1) рівномірно розподілятися по ГС [34];

2) мати достатньо великий обсяг, якого вистачає для правильних висновків про ГС [35].

Розрізняють два типи однорідності вибірки: лінгвістична та статистична.

У межах лінгвістичної однорідності вибірки виділяють [3]:

– хронологічну (тексти вибірки повинні мати хронологічні межі) [36];

– жанрову (тексти вибірки повинні бути жанрово обмежені) [37];

– тематичну (тексти повинні бути тематично обмежені) [38].

Статистично однорідною є вибірка, в якій досліджувані одиниці мають статистичну поведінку, яка суттєво між собою не відрізняється [39]. Якщо середня частота явища (літери, морфеми, слова, довжини слова, довжини речення і т. д.) в одній вибірці суттєво не відрізняється від його частоти в інших вибірках, то ці вибірки статистично однорідні стосовно цього явища [40].

За способом організації виділяють такі різновиди вибірок [3]:

– механічна – організована з урахуванням рівномірності розподілу досліджуваної одиниці по генеральній сукупності [41]. Всі тексти генеральної суку-

пності перенумеровують, а потім, наприклад, з кожного п'ятого, десятого, двадцятого тексту вибирають відрізок необхідної довжини [42].

– випадкова – організована шляхом випадкового вибору текстів з ГС [43]. В основі такого методу організації вибірки лежить гіпотеза про те, що досить велика кількість навздогад відібраних одиниць з ГС повинна адекватно її представляти [44]. Тож кожна сторінка, розділ чи інша одиниця тексту ГС повинні мати однаковий шанс потрапити до вибірки. Тому, як правило, випадкова вибірка ґрунтується на таблиці випадкових чисел [45].

– зональна (типова) – організована на основі лінгвістично однорідної сукупності текстів, тобто зони [46]. Зоною залежно від мети дослідження вважають прозу, поезію та драму в художній літературі; твори одного автора або конкретний твір; сукупність слів певної морфемної структури (наприклад, префіксальних або одноморфемних) тощо [47].

Вибірка може бути структурною, тобто складатися із менших частин (підвибірками) та неструктурною, тобто суцільною [48].

Співвідношення між частотою одиниць мови та мовлення покажемо на прикладі: «якщо взяти з лото 33 бочечки, розклеїти на них український алфавіт і перемішати, то ймовірність того, що перша витягнута бочечка виявиться із чистою голосною літерою, буде 6:33 (6 чистих голосних букв (а, о, у, є, и, і) до 33 усіх букв українського алфавіту), тобто приблизно 16 %» [3]. Якщо ж узяти випадковий український текст і вибрати з нього навгад одну літеру, то ймовірність того, що виявиться чистою голосною буде приблизно 30 % [3]. У першому випадку йдеться про ймовірність групи з шести літер на рівні парадигматики (мови), у другому – на рівні синтагматики (мовлення) [49]. Припустити, що всі голосні звуки або всі відмінкові форми, або всі члени речення рівноймовірні, означало би підмінити природне мовлення його схемою [50]. Отже, мовлення надає перевагу невеликій кількості одиниць (закон переваги), які й становлять ядро мовленнєвої підсистеми, тоді як у мові всі одиниці рівноймовірні [51]. У різних мовах частота тієї самої букви чи послідовності букв неоднакова, тому, знаючи порядок найчастотніших букв, біграмів, триграмів, чотириграмів певної мови, можна автоматично її ідентифікувати [52]. Частотність цих одиниць у мові визначають на репрезентативних вибірках, оскільки у творах конкретних авторів, стилів чи тем частотність також різна [53]. Наприклад, для українських текстів виявлено, що статистичними параметрами стилів можна вважати частоти голосних, приголосних, пропуски між словами, а також груп приголосних: м'яких, сонорних [1–7]. Частоту букв у текстах досліджували для потреб криптографії (науки про зашифровування та розшифрування повідомлень), зокрема, азбуки морзе (чим частотніша літера чи буквосполучення, тим коротші риски для їхнього позначення), для стенографування, автоматичного визначення мови, підтвердження чи заперечення авторства твору тощо [54]. Морфеми та граматичні категорії також мають власні кількісні характеристики:

– неоднорідне використання морфем іншомовного походження та питомо мовних [55];

– дієслів теперішнього, минулого та майбутнього часу, дійсного, умовного, наказового способу [56];

– форм дієслова (інфінітива, особових форм, дієприкметника, дієприслівника, безособових форм на -но, -то) [57];

– різних частин мови залежно від стилю [58].

Виявлено закономірність, що в різних функціональних стилях кількісне співвідношення функціонування різних відмінків неоднакове [59]. Наприклад, наукова проза надає перевагу родовому і нехтує називним відмінком, а розмовне мовлення навпаки і т. д. [3].

Кількісні характеристики слів найкраще видно з ЧС [59]. Функціональна залежність зв'язку між частотністю слова та полісемією, а також між частотністю слова та його рангом у словнику за спадом частот виражає закон Ціпфа-Мандельброта [3]. Найчастотнішими є службові частини мови або загальні абстрактні поняття [60]. Натомість слова з конкретним значенням (необхідні для розмови у звичайній ситуації) — низькочастотні [61]. Хоча вживаються рідко, проте завжди є у свідомості мовця [62]. Іншими словами, критерій частотності доповнено критерієм тематичності [3]. Формулу для встановлення ступеня синонімічності (семантичної близькості) слів [63]: $C=2c/(n_1+n_2)$, де n_1 – кількість значень першого слова, n_2 – кількість значень другого слова, c – кількість спільних значень у даної пари слів. Кількісні характеристики синтаксичних конструкцій теж залежать від функціонального стилю: в розмовно-побутовому переважають прості неускладнені, навіть неповні та обірвані речення, в науковому та офіційно-діловому – складні речення, ускладнені зворотами, вставними і вставленими конструкціями [64].

Темп мовлення-думки спрощено можна подати відношенням кількості самостійних слів до кількості простих речень, оскільки чим менше слів входить до одного речення, тим частотніші речення (а, значить, і думки) [65]. Виявлено, що темп мовлення-думки в казці – 2,39, а в науковому тексті – всього 0,42 [3]. Це означає, що мовлення і дія у казці розгортається швидше майже у 6 разів. І це зрозуміло: у казці думки та висловлювання, якими виражені, прості за структурою, тому й швидші, легше вибудовуються в динамічну послідовність; у науковій статті структура думки-мовлення набагато складніша, тому канали свідомості пропускають одиниці такого мовлення-думки повільніше [66].

Коефіцієнт зв'язності мовлення логічно вимірювати, взявши за основу відношення кількості прийменників та сполучників до кількості окремих речень [67]. Нехай цей коефіцієнт дорівнює одиниці тоді, коли в одному реченні є три сполучні елементи (прийменники та сполучники) [3]: $K_z=(P+C)/(3N)$, де P – кількість прийменників, C – кількість сполучників, N – кількість окремих речень. Виявлено, що текст казки має коефіцієнт зв'язності 0,77, а текст наукової статті – 3,0, тобто зв'язність у другому тексті у 3,9 разів сильніша, ніж у першому [3].

Поняття індекс синтетичності мови визначаємо як M/W, де M – кількість морфів у певному відрізку тексту, W – кількість слів у цьому тексті [68]. Мови з індексом від 1 до 2 вважаються аналітичними, від 2 до 3 – синтетичними, а від 3 і більше – полісинтетичними [69]. Найнижчу величину має в'єтнамська мова – 1,06, тобто на 100 слів припадає 106 морфів, найвищу має ескімоська мова – 3,72, тобто на 100 слів припадає 372 морфи [3]. Англійська мова має показник 1,68, російська – 2,33... [3]. На підставі індексу синтетичності до аналітичних

мов відносять в'єтнамську, китайську, перську, італійську, німецьку, данську; до синтетичних – українську, російську, санскрит, литовську, чеську, польську, якутську: до полісинтетичних – ескімоську, тубільно-американські, іберокавказькі [69]. У зв'язку зі збільшенням доступності та поширення текстових документів в електронній формі збільшилася на важливості використання автоматичних методів для аналізу змісту документів [70]. До завдань аналізу тексту можна віднести завдання класифікації і кластеризації документів за різними критеріями, наприклад, жанру, епосі написання, формату (роман, есе, нарис), емоційному забарвленню, стилю мовлення, а також завдання визначення автора тексту [71]. З спрощенням доступу до різними даними, розширенням можливості пошуку, копіювання та розповсюдження даних в мережах стає актуальною задача ідентифікації автора [72]. Так само, питання, пов'язані з встановленням авторства, є важливими в лінгвістичних, історичних і криміналістичних дослідженнях [73]. Загальнодоступність електронних пристроїв дозволяє відсунути розпізнавання автора з залученням великої кількості експертів на другий план, прискорити і спростити цей процес за допомогою його автоматизації [74]. Поняття ідентифікація автора визначається як процес встановлення автора по безлічі загальних і приватних ознак тексту, що складають авторський стиль [75].

В існуючих системах визначення авторства тексту користуються популярністю статистичні методи, засновані на пошуку «авторського інваріанта» [76]. «Авторський інваріант» характеризує мовну особливість (лексичну, граматичну, фразеологічні та іншу) тексту [77]. Як інваріанта можуть виступати: частка голосних або приголосних, частота вживання певної частини мови, ймовірність переходів від однієї частини мови до іншої, «улюблені» слова, інформаційна ентропія і так далі [78]. В [3] запропоновано статистичний метод визначення автора і жанру тексту, заснований на розподілі частот буквосполучень (*n*-грам). Метод показав достойні результати для творів російської літератури [79]. Але точність статистичних методів визначення авторства сильно залежить від специфіки використовуваних даних: від мови, на якому написані тексти [80], від стилю мовлення тексту [82], і, перш за все, від довжин текстів, на яких проводять дослідження [83]. В силу цього важко робити висновок про точність такого підходу на даних іншої природи. З цієї причини метою цієї роботи був аналіз застосовності такого математичного апарату, як розподіл частот буквосполучень для різних мов при вирішенні завдання встановлення авторства текстів, що мають різні довжини і написаних в різному стилі мовлення [84].

В якості критерію близькості двох текстів обчислюють відстань між відповідними векторами [85]. Набори параметрів та коефіцієнти мовлення подають як звичайні вектори в *n*-вимірному декартовому просторі з початку координат [86]. Тоді відстанню між текстами є звичайна декартова відстань між кінцями відповідних векторів. Така нормова відстань є інтегральною характеристикою відмінності текстів [87]. І тексти з великою відстанню з високою ймовірністю належать різним авторам. Таким чином, щоб співставити авторство двох текстів, досить обчислити для них параметри і визначити відстань [88]. Щоб зіставити текст з автором, порівнюються вектори параметрів автора і даного тексту, тобто фактично знову порівнюють два тексти – текст зі свідомо відомим авто-

ром (еталонний текст) і текст, авторство якого потрібно встановити, підтвердити або спростувати (аналізований/досліджуваний текст) [89]. Складають також вектори формальних параметрів, що розрізняють не конкретних авторів (або групи), а виділяють певні характеристики авторів (наприклад, освітній рівень) [90]. У більшості випадків згідно [91] в якості характеристичних параметрів тексту обирають статистичні характеристики:

– кількість використання певних частин мови, деяких конкретних слів, знаків пунктуації, фразеологізмів, архаїзмів, рідкісних та іноземних слів [92];

– кількість і довжина речень (виміряна в словах, складах, знаках), середня довжина речення [93];

– кількість повнозначних і службових слів [94];

– обсяг словника, відношення кількості дієслів до загальної кількості слововживань в тексті тощо [95].

Основна проблема формальних методів аналізу авторства полягає якраз у виборі параметрів та коефіцієнтів мовлення [96]. Існує цілий ряд формальних статистичних характеристик текстів, непридатних для визначення авторства в силу одного з двох недоліків [1–7, 97]:

– Відсутність стійкості. Розкид значень параметра для текстів одного і того ж автора настільки великий, що діапазони можливих значень для різних авторів перетинаються [99]. Очевидно, даний параметр не допоможе розрізняти авторів, а при використанні в складі групи параметрів лише зіграє роль додаткового інформаційного шуму [100].

– Відсутність здібності розрізняти. Параметр може приймати близькі значення для всіх або більшості авторів, оскільки його значення визначають властивостями мови, на якому написані тексти, а не індивідуальними особливостями автора тексту [101]. Тому параметри повинні попередньо досліджуватися на стійкість і здатність розрізняти, бажано на текстах великої кількості різних авторів [102].

В роботах [1–7] виділені наступні умови застосовності формального коефіцієнта мовлення стилю автора:

– Масовість (використання тих характеристик тексту, які слабо контролюються автором на свідомому рівні, щоб усунути можливість свідомого спотворення автором характерного для нього стилю або імітації стилю іншого автора) [103].

– Стійкість (збереження постійно значення для одного учасника, але деяке відхилення значень від середнього має бути досить малим) [104].

– Здатність розрізняти (приймає істотно різні значення для різних авторів, тобто перевищують коливання, можливі для одного учасника) [105].

Обрати коефіцієнти та параметри мовлення, які гарантовано розрізняють двох будь-яких авторів, дуже важко [106]. Якими б не були параметри, завжди існує ймовірність того, що два або більше учасника є за даними параметрами близькими в силу випадкового збігу [107]. Тому на практиці є достатнім, щоб параметр дозволяв впевнено розрізняти між собою різні підмножини авторів, тобто існувала б досить велика кількість підмножин авторів, для яких середні значення параметра значно відрізняються [108]. Параметр, очевидно, не допоможе розрізнити тексти авторів з однієї підмножини, але дозволить впевнено розрізняти тексти авторів, які потрапили в різні підмножини [1–7]. Розрізняти тексти авторів однієї підмножини

можна за рахунок використання одночасно досить великого вектора різних за характером параметрів – в цьому випадку ймовірність випадкового збігу стане помітно меншою. Для впевненого виведення текстів, для яких формально обчислена параметрична відстань мала, необхідно провести додаткове дослідження експертними методами, наприклад, аналіз ключових і/або стопових (службових) слів [1–7].

Отже, виникає необхідність із-за відсутності практичних експериментів визначення стилю автора для україномовних науково-технічних текстів провести дослідження в цьому напряму. Для розв'язку задачі плагіату як копірайту в наш час вже розроблено багато систем. Щодо рерайту – для слов'янських мов досить складно вирішити таку задачу із наявності великої множини синонімів та можливості перебудови речень з використанням інших закінчень. Це питання не стосується використання службових слів, так як більшість людей при плагіаті навіть не звертає на них увагу. Тому це і спонукає досліджувати задачу ідентифікації стилю автора для визначення ступеню належності конкретного тексту конкретному автору.

3. Мета і завдання дослідження

Метою дослідження є розроблення методу автоматичного визначення авторства текстового контенту на основі статистичного аналізу розподілу N-грам.

Для досягнення мети були поставлені такі завдання:

– розробити квантитативний метод ідентифікації потенційного автора тексту із множини можливих на основі порівняння результатів аналізу еталонного тексту з досліджуваним;

– розробити програмне забезпечення контент-моніторінгу для визначення автора в україномовних текстах на основі лінгвостатистичного аналізу еталонного текстового контенту;

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

4. Квантитативний метод

Квантитативний метод ідентифікації потенційного автора тексту із множини можливих на основі порівняння результатів аналізу еталонного тексту з досліджуваним

Лінгвометрія є галузю прикладної лінгвістики, що виявляє, вимірює та аналізує кількісні характеристики одиниць різних рівнів мови чи мовлення [3]. Застосовуючи апарат математичної статистики, лінгвометрія бере участь у вирішенні таких завдань мовознавства, як створення:

– словників (у тому числі частотних і статистичних) та порівняння,

– автоматичних словників, тезаурусів,

– систем стенографії,

– методів та засобів автоматичного визначення мови,

– методів та засобів інформаційного пошуку тощо.

Кожна мова має власні статистичні параметри, і знання частоти поява літер та їх сполучень (2-грам, 3-грам, 4-грам) певної мови дає змогу автоматично її іден-

тифікувати. Наприклад, для українських текстів було виявлено, що статистичними параметрами стилів можна вважати частоти голосних, приголосних, пропуски між словами, а також м'яких і сонорних груп приголосних [3]. Покажемо, як виконати оцінювання мовлення конкретного автора на конкретному уривку його роботи [77] за допомогою певного еталону – наприклад, даних про частотність літер української мови. Розглянемо два уривки технічного тексту українською мовою, подані у форматі, де літери розташовані за спаданням частот їх появи в уривку (частоти подано у табл. 1), а розрізнення на малу та велику літери не здійснюється. Знайдемо тип кореляції частот літер уривків [76] та еталону [77], результати, що підтверджують висновки, подамо, зокрема, у графічному вигляді.

У табл. 1 для зручності внесено такі дані: частотність вживання літер української мови, абсолютні та відносні частоти вживання літер у досліджуваних Уривку 1 Автора 1 [76] та Уривку 2 Автора 2[77]. Зауважимо, що Уривок 1 містить 556 символів, Уривок 2 містить 541 символ. Відзначимо, що поняття «інші» у стовпці літер містить автентичні літери для української мови (ї, є, ґ, і), що є маловживані в більшості технічних текстах. Це дає змогу досягти певної незалежності під час аналізу. Зобразимо на рис. 1 графічно отримані результати.

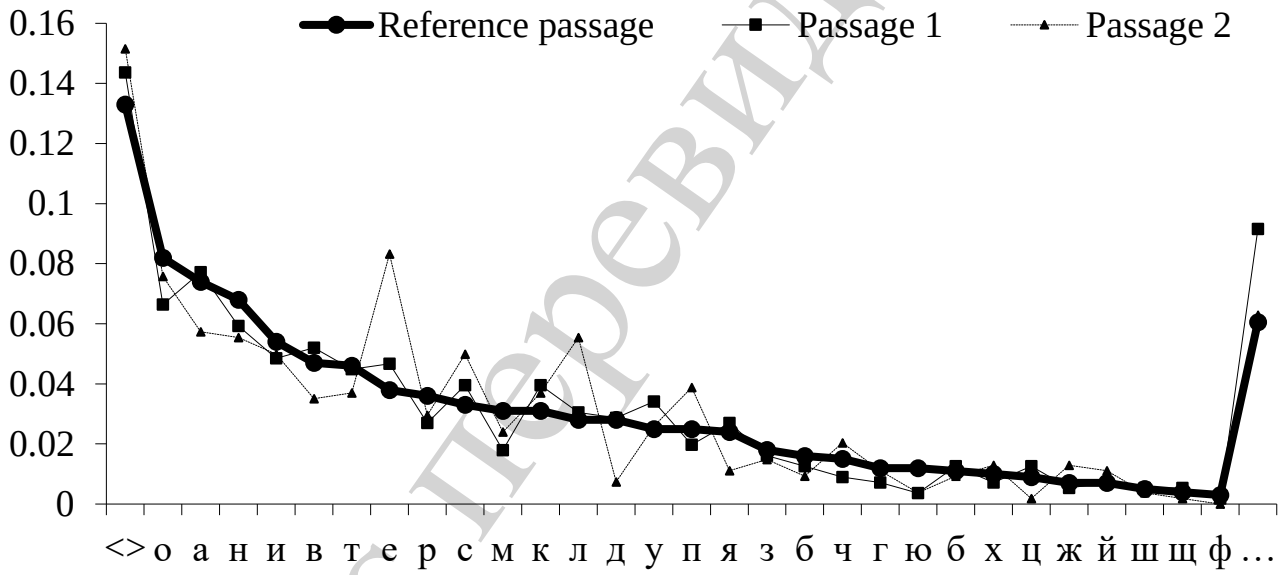


Рис. 1. Графічне подання відносних частот появи літер в еталоні та досліджуваних уривках

Графічне подання відносних частот появи літер в уривках дає переконливої відповіді на запитання, який із уривків написаний яким автором.

Розподіл 1-грам у роботах різний. Оптимальними показниками дослідження текстів є аналіз 3-грам [3]. Це перевіримо в наступних етапах дослідження. Є різкий стрибок відносної частоти появи літери «е» для Уривку 2 відносно еталонних значень Еталону 1 [77] (рис. 2), тому вважатимемо, що з більшою ймовірністю Еталон 1 написаний автором Уривку 1 [76]. Наведемо також чисельні значення кореляції частотності літер в уривках та еталоні. Знайдемо два коефіцієнти кореляції: для еталону та Уривку 1 [76] і для еталону та Уривку 2 [78]; ближчий до 1

коефіцієнт свідчитиме про більшу ймовірність належності відповідного уривку до еталону. Обчислення коефіцієнту кореляції для еталону та Уривку 1 дають $R_{е\text{-}у1}$=0,962716, а коефіцієнту кореляції для еталону та Уривку 2 – $R_{е\text{-}у2}$=0,909958. Аналогічно, значення відносних частот в Еталоні 2 та Уривках 1, 2 на рис. 3 суттєво різняться, тому ймовірнісно що автор Еталону 2 [75] не є автором уривків 1, 2.

Таблиця 1

Частотності появи літер в еталоні та досліджуваних уривках

Літера	Частотність вживання літер української мови (еталон)	Абсолютна частота літер в Уривку 1	Відносна частота вживання літер в Уривку 1	Абсолютна частота літер в Уривку 2	Відносна частота вживання літер в Уривку 2
« »	0,133	80	0,14	82	0,15
о	0,082	37	0,07	41	0,08
а	0,074	43	0,08	31	0,06
н	0,068	33	0,06	30	0,06
и	0,054	27	0,05	27	0,05
в	0,047	29	0,05	19	0,04
т	0,046	25	0,04	20	0,04
е	0,038	26	0,05	45	0,08
р	0,036	15	0,03	16	0,03
с	0,033	22	0,04	27	0,05
м	0,031	10	0,02	13	0,02
к	0,031	22	0,04	20	0,04
л	0,028	17	0,03	30	0,06
д	0,028	16	0,03	4	0,01
у	0,025	19	0,03	14	0,03
п	0,025	11	0,02	21	0,04
я	0,024	15	0,03	6	0,01
з	0,018	9	0,02	8	0,01
б	0,016	7	0,01	5	0,01
ч	0,015	5	0,01	11	0,02
г	0,012	4	0,01	6	0,01
ю	0,012	2	0,00	2	0,00
б	0,011	7	0,01	5	0,01
х	0,01	4	0,01	7	0,01
ц	0,009	7	0,01	1	0,00
ж	0,007	3	0,01	7	0,01
й	0,007	4	0,01	6	0,01
ш	0,005	3	0,01	2	0,00
щ	0,004	3	0,01	1	0,00
ф	0,003	1	0,00	0	0,00
інші	0,0605	51	0,09	34	0,06

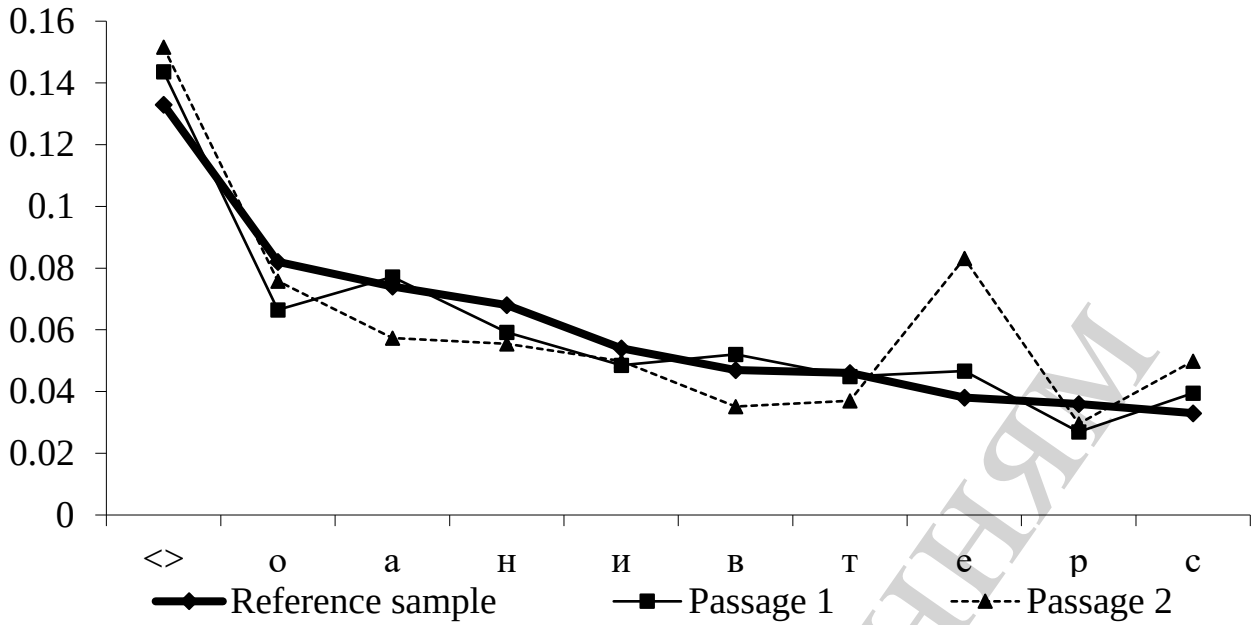


Рис. 2. Графічне подання відносних частот появи десятьох найбільш частотних символів у Еталоні 1 та досліджуваних Уривках 1, 2, включно із пропуском

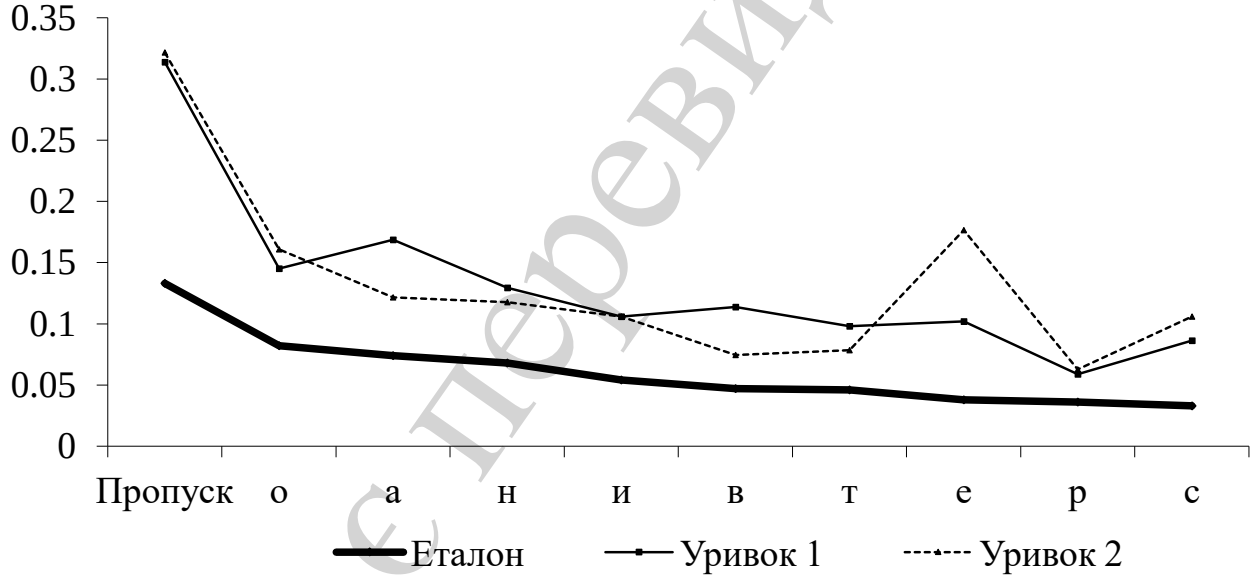


Рис. 3. Графічне подання відносних частот появи десятьох найбільш частотних символів у Еталоні 2 та досліджуваних Уривках 1, 2, включно із пропуском

Отримані значення коефіцієнтів, а також аналіз графічних результатів дає змогу стверджувати, що ймовірність належності Уривку 1 [76] до Еталону 1 [77] вища, ніж для Уривку 2 [75].

5. Програмне забезпечення контент-моніторінгу для визначення автора в україномовних текстах

Для досягнення мети дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі Victana

[26]. Для якісного та ефективного аналізу контенту при визначенні ступеня авторства конкретної людини пропонуємо аналізувати еталонного тексту та досліджуваного в декілька етапів:

– лінгвометричний аналіз коефіцієнтів різноманіття авторського мовлення (алг. 1);

– стилометричний аналіз (алг. 2);

– аналіз стійких словосполучень (алг. 3);

– лінгвостатистичний аналіз через N-грам (алг. 4).

На Web-ресурсі для лінгвометричного аналізу є такі поля (рис. 4):

– Знаків. (Введений текст повинен містити не менше 100 та не більше 10000 знаків.) – виставляється максимальний розмір контенту.

– Контент – поле, куди копіюється із буфера досліджуваний текст.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

Рис. 4. Приклад результату застосування лінгвометричного аналізу

Алгоритм 1. Лінгвометричний аналіз тексту для визначення авторства.

Крок. 1. Перевірка довжини тексту – лишнє відсікається.

Крок. 2. Визначення кількості речень.

Крок. 3. Очищення досліджуваного тексту (цифри, спецсимволи).

Крок. 4. Визначення загальної кількості слів у тексті *N*.

Крок. 5. Визначення кількості слів *W*.

Крок. 6. Визначення кількості прийменників *Z*.

Крок. 7. Визначення кількості сполучників *S*.

Крок. 8. Розрахунок коефіцієнтів авторського мовлення.

Крок. 9. Вивід результатів кінцевому користувачу (табл. 2, рис. 4).

Таблиця 2

Приклад розрахунків коефіцієнтів авторського мовлення

№	Коефіцієнт	Вхідні дані	Розрахунок
1.	Коефіцієнт лексичної різноманітності: $K_l=W/N$	W=184 N=295	K_l=0.62372881355932
2.	Коефіцієнт синтаксичної складності: $K_s=1–P/W$	P=18 W=184	K_s=0.90217391304348
3.	Коефіцієнт зв'язності мовлення: $K_z=(Z+S)/(3*P)$	Z=20 S=28 P=18	K_z=0.88888888888889
4.	Індекс винятковості: $I_{wt}=W_1/W$	W_1=141 W=184	I_{wt}=0.76630434782609
5.	Індекс концентрації: $I_{kt}=W_{10}/W$	W_{10}=2 W=184	I_{kt}=0.010869565217391

На Web-ресурсі для стилеметричого аналізу є такі поля (рис. 5):

– Еталонний текст – поле, куди копіюється із буфера Еталонний текст.

– Вибрати Уривок 1 (2, 3) – відкриваємо доступ до уривків. Доступ до наступного уривку тільки після активації доступу до попереднього. Доступ відкривається послідовно від меншого числа до більшого.

– Уривок 1 (2, 3) – поле, куди копіюється із буфера текст відповідного уривку.

Введений текст повинен містити не менше 100 знаків. (Зараз 0) – Після запуску розрахунку буде розраховано та показано реальну кількість знаків кожного уривку окремо.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

Алгоритм 2. Стилеметричний аналіз тексту для визначення авторства.

Крок. 1. Перевірка довжин еталонного тексту та вибраних уривків та приведення довжини еталонного тексту до мінімального із перевірених.

Крок. 2. Очищення еталонного тексту від спецсимволів та інш.

Крок. 3. Визначення кількості слів у тексті еталону.

Крок. 4. Визначення кількості стоп-слів (прийменників + сполучників + часток) у тексті еталону (рис. 6, 7).

Крок. 5. Довжина Уривка 1 не більше мінімально тексту.

Крок. 6. Очищення Уривка 1 від спецсимволів та інш.

Крок. 7. Визначення кількості слів W1 для Уривка 1.

Крок. 8. Визначення кількості стоп-слів (прийменників + сполучників + часток) в тексті.

Крок. 9. Підготовка окремих масивів (уривок та еталон) для розрахунку коефіцієнта кореляції (рис. 7).

Крок. 10. Виклик функції для розрахунку коефіцієнта кореляції.

Крок. 11. Формування масиву для формування графічного зображення відносної частоти появи стопових слів в Уривку 1 та в еталоні.

Крок. 12. Виклик функції для розрахунку графіка ВЧ (рис. 8).

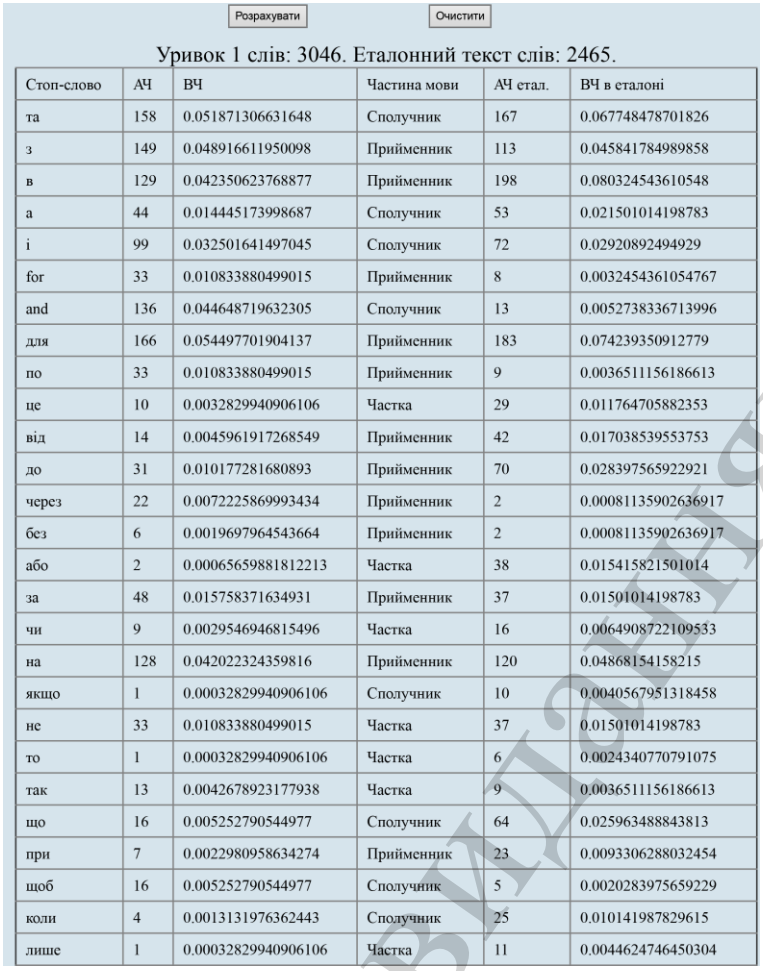
Крок. 13. Виклик функції для розрахунку коефіцієнта кореляції Уривків 2(3) для кожного зі службових слів.

Крок. 14. Формуємо слова списку Сводеша із довідника, визначення кількості слів із списку Сводеша в тексті уривку (для еталонного тексту та вибраних уривків – табл. 3).

Крок. 15. Формуємо спільні для Еталону, Уривків 1–3 та списку Сводеша.

Крок. 16. Результати дослідження виводяться на екран (табл. 4)

Рис. 5. Приклад введення даних для стилеметричого аналізу



Розрахувати Очистити

Уривок 1 слів: 3046. Еталонний текст слів: 2465.

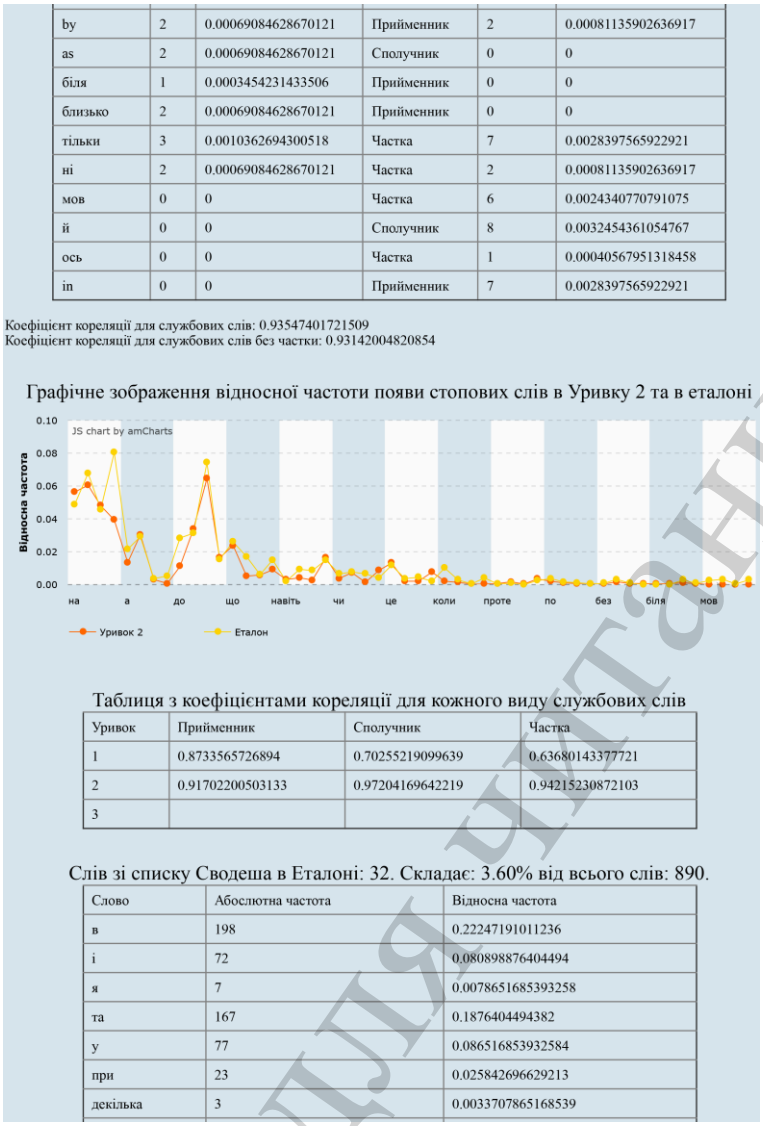
Стоп-слово	АЧ	ВЧ	Частина мови	АЧ етал.	ВЧ в еталоні
та	158	0.051871306631648	Сполучник	167	0.067748478701826
з	149	0.048916611950098	Прийменник	113	0.045841784989858
в	129	0.042350623768877	Прийменник	198	0.080324543610548
а	44	0.014445173998687	Сполучник	53	0.021501014198783
і	99	0.032501641497045	Сполучник	72	0.02920892494929
for	33	0.010833880499015	Прийменник	8	0.0032454361054767
and	136	0.044648719632305	Сполучник	13	0.0052738336713996
для	166	0.054497701904137	Прийменник	183	0.074239350912779
по	33	0.010833880499015	Прийменник	9	0.0036511156186613
це	10	0.0032829940906106	Частка	29	0.011764705882353
від	14	0.0045961917268549	Прийменник	42	0.017038539553753
до	31	0.010177281680893	Прийменник	70	0.028397565922921
через	22	0.0072225869993434	Прийменник	2	0.00081135902636917
без	6	0.0019697964543664	Прийменник	2	0.00081135902636917
або	2	0.00065659881812213	Частка	38	0.015415821501014
за	48	0.015758371634931	Прийменник	37	0.01501014198783
чи	9	0.0029546946815496	Частка	16	0.0064908722109533
на	128	0.042022324359816	Прийменник	120	0.04868154158215
якщо	1	0.00032829940906106	Сполучник	10	0.0040567951318458
не	33	0.010833880499015	Частка	37	0.01501014198783
то	1	0.00032829940906106	Частка	6	0.0024340770791075
так	13	0.0042678923177938	Частка	9	0.0036511156186613
що	16	0.005252790544977	Сполучник	64	0.025963488843813
при	7	0.0022980958634274	Прийменник	23	0.0093306288032454
щоб	16	0.005252790544977	Сполучник	5	0.0020283975659229
коли	4	0.0013131976362443	Сполучник	25	0.010141987829615
лише	1	0.00032829940906106	Частка	11	0.0044624746450304

Рис. 6. Приклад результату застосування стилеметричого аналізу

Таблиця 3

Уривок 1 слів: 153. Еталонний текст слів: 153

Слово	АЧ	ВЧ	Частина мови	АЧ етал	ВЧ в еталоні
в	5	0.032679738562	Прийменник	5	0.032679738562
а	2	0.0130718954248	Сполучник	2	0.0130718954248
це	1	0.0065359477124	Частка	1	0.0065359477124
та	16	0.1045751633987	Сполучник	16	0.1045751633987
для	7	0.0457516339869	Прийменник	7	0.0457516339869
з	2	0.0130718954248	Прийменник	2	0.0130718954248
ж	1	0.0065359477124	Частка	1	0.0065359477124
і	3	0.019607843137	Сполучник	3	0.019607843137
також	2	0.0130718954248	Сполучник	2	0.0130718954248
мов	2	0.0130718954248	Частка	2	0.0130718954248
у	1	0.0065359477124	Прийменник	1	0.0065359477124
що	1	0.0065359477124	Сполучник	1	0.0065359477124
за	1	0.0065359477124	Прийменник	1	0.0065359477124



by	2	0.00069084628670121	Прийменник	2	0.00081135902636917
as	2	0.00069084628670121	Сполучник	0	0
біля	1	0.0003454231433506	Прийменник	0	0
близько	2	0.00069084628670121	Прийменник	0	0
тільки	3	0.0010362694300518	Частка	7	0.0028397565922921
ні	2	0.00069084628670121	Частка	2	0.00081135902636917
мов	0	0	Частка	6	0.0024340770791075
й	0	0	Сполучник	8	0.0032454361054767
ось	0	0	Частка	1	0.00040567951318458
in	0	0	Прийменник	7	0.0028397565922921

Таблиця з коефіцієнтами кореляції для кожного виду службових слів

Уривок	Прийменник	Сполучник	Частка
1	0.8733565726894	0.70255219099639	0.63680143377721
2	0.91702200503133	0.97204169642219	0.94215230872103
3			

Слів зі списку Сводеша в Еталоні: 32. Складає: 3.60% від всього слів: 890.

Слово	Абослютна частота	Відносна частота
в	198	0.22247191011236
і	72	0.080898876404494
я	7	0.0078651685393258
та	167	0.1876404494382
у	77	0.086516853932584
при	23	0.025842696629213
декілька	3	0.0033707865168539

Рис. 7. Приклад результату застосування стилеметричого аналізу для Уривку 2

Таблиця 4

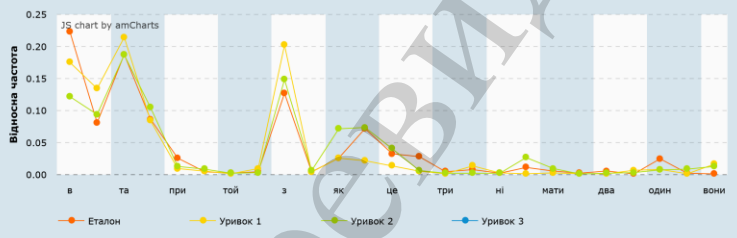
Слова, спільні для Еталону, Уривків 1–3 та списку Сводеша: 8. Складає: 26.67 % від всього слів: 30

№	Спільні	АЧ	Еталон	Уривок 1	Уривок 2	Уривок 3
1	в	5	0.167	0.167	0.167	0.167
2	це	1	0.033	0.033	0.033	0.033
3	та	16	0.533	0.533	0.533	0.533
4	з	2	0.167	0.167	0.167	0.167
5	коло	1	0.033	0.033	0.033	0.033
6	і	3	0.1	0.1	0.1	0.1
7	у	1	0.033	0.033	0.033	0.033
8	що	1	0.033	0.033	0.033	0.033

Для автоматизованого опрацювання тексту має велике значення не тільки те, яка частота появи тієї чи іншої категорії, а взагалі її присутність в досліджуваному тексті. Кількісний підрахунок дозволяє зробити об’єктивні висновки

щодо спрямованості матеріалів за кількістю уживань одиниць аналізу (ключових цитат) в досліджуваних текстах. Якісний аналіз робить те саме, але внаслідок вивчення того, чи зустрічається (і в якому контексті) якась важлива, оригінальна категорія взагалі. Підводячи підсумки, слід зазначити, що використання контент-аналізу для створення інформаційних систем дозволяє вловити поширеність тієї чи іншої ознаки досліджуваної сукупності текстів. При цьому важливо не стільки абсолютне, скільки відносне значення ознаки, тобто характеристика її місця (частки) серед інших ознак. Вимірювання співвідношення між ознаками в текстах дає емпіричний матеріал для розуміння функціональних зв'язків між елементами відображеної в текстах дійсності. При наявності текстів, що мають хронологічну послідовність, можна мати низку фіксованих у часі портретів досліджуваної дійсності, що дає змогу висувати гіпотези прогностичного характеру про функціонування елементів системи. Наприклад, частотні характеристики тексту (середній розмір речень) може свідчити про певну специфіку інтелектуальних здібностей особи у плані вербального подання думок. За визначенням середнього розміру речень можна дати характеристику зміни емоційного стану індивіда.

Графічне зображення ВЧ появи слів, спільних для Уривків 1–3 з еталоном та списку Сводеша



Слова, спільні для Еталону, Уривків 1–3 та списку Сводеша

№ зп	Спільні слова	ВЧ в Еталоні	ВЧ в Уривку 1	ВЧ в Уривку 2	ВЧ в Уривку 3
1	в	0.22247191011236	0.17503392130258	0.12166488794023	
2	і	0.080898876404494	0.13432835820896	0.093916755602988	
3	та	0.1876404494382	0.21438263229308	0.18676627534685	
4	у	0.086516853932584	0.084124830393487	0.10458911419424	
5	при	0.025842696629213	0.0094979647218453	0.012806830309498	
6	те	0.0044943820224719	0.005427408412483	0.0085378868729989	
7	той	0.002247191011236	0.001356852103120 8	0.0021344717182497	
8	зі	0.0044943820224719	0.0094979647218453	0.0032017075773746	
9	з	0.12696629213483	0.20217096336499	0.14834578441836	
10	цей	0.0044943820224719	0.0040705563093623	0.0064034151547492	
11	як	0.025842696629213	0.025780189959294	0.071504802561366	
12	що	0.071910112359551	0.021709633649932	0.073639274279616	
13	це	0.032584269662921	0.013568521031208	0.040554962646745	
14	коли	0.028089887640449	0.005427408412483	0.0064034151547492	
15	три	0.0044943820224719	0.001356852103120 8	0.0021344717182497	
16	де	0.0078651685393258	0.013568521031208	0.0021344717182497	
17	ні	0.002247191011236	0.0027137042062415	0.0021344717182497	
18	якщо	0.01123595505618	0.001356852103120 8	0.026680896478122	
19	мати	0.0056179775280899	0.0027137042062415	0.0096051227321238	
20	великий	0.002247191011236	0.001356852103120 8	0.0010672358591249	
21	два	0.0056179775280899	0.001356852103120 8	0.0021344717182497	
22	багато	0.001123595505618	0.0067842605156038	0.0032017075773746	

Рис. 8. Приклад результату стилеметричого аналізу для Уривків 1–3

Одним із найбільш значних та найпотужніших у психолінгвістичній діагностиці тексту є вибір аналізу словникового варіанта в контекстній залежності. Завдяки встановленню коефіцієнта словникової різноманітності (табл. 5) в мовленні людини, можна виявити психопатологію, наприклад, шизофренію, а також схильність до неї.

Таблиця 5

Коефіцієнти частотних характеристик тексту

Коефіцієнт	Формула
Словникової різноманітності	$К_{слов.різном.}$=різних слів/$2N_{всіх\ слів}$
Дієслівності (агресивності)	$К_{дієсл.}$=дієслів/$N_{всіх\ слів}$·100 %
Емоційності тексту	$К_{прикм.}$=прикм/$2N_{всіх\ слів}$
Логічної зв'язності	$К_{лог.зв'язн.}$=служб. слів/$3N_{реч}$
Емболії (засміченості)	$К_{емб.}$=ембол/$N_{всіх\ слів}$·100 %

Іншим критерієм мовної компетенції є коефіцієнт дієслівності (агресивності). Суть цього коефіцієнту полягає у співвідношенні кількості дієслів і дієслівних форм (дієприкметників і дієприслівників) до загальної кількості всіх слів. Як і в психології, високий коефіцієнт агресивності свідчить про можливу високу емоційну напруженість автора, яка відображена в самому тексті проявами зміни динаміки подій та іншими характерними особливостями. Коефіцієнт логічної зв'язності також вираховують за формулою співвідношення загальної кількості службових слів (сполучників, прийменників і часток) до загальної кількості речень. При величинах у межах одиниці забезпечується достатньо гармонійний зв'язок службових слів і синтаксичних конструкцій. Коефіцієнт емболії (мед. ембола – закупорка кровоносної судини), чи «засміченості» мовлення – це співвідношення загальної кількості ембол (слів, які не несуть семантичного навантаження) до загальної кількості слів у реченні. До складу ембол входять вигуки (ну-ну, ха-ха, еге, ж, ой тощо), вульгаризми (ненормативна лексика), непотрібні повторення. Коефіцієнт емболії свідчить про особливості вербального інтелекту й емоційний стан мовця/ автора тексту. Також може дати уявлення про загальну культуру мовлення. Навіть враховуючи той факт, що художній текст в принципі вважається андрогенним та є переплетінням функцій підрядності – якостей авторського «Я», певним чином градуюються в залежності від характерологічного профілю того чи іншого автора. Іншими словами, текст оригіналу і текст перекладу знаходяться у залежності від їх авторів.

На Web-ресурсі для аналізу стійких словосполучень є такі поля (рис. 9):

– Введіть кількість словосполучень для виведення на екран (10;100) – скільки словосполучень буде виведено на екран після розрахунку.

– Вибрати Уривок 1 (2, 3) – відкриваємо доступ до уривків. Доступ до наступного уривку тільки після активації доступу до попереднього. Доступ відкривається послідовно від меншого числа до більшого. (Не реалізовано – аналізується тільки один уривок)

– Уривок 1 – поле, куди копіюється із буфера текст відповідного уривку.

– Використано:57 % Введений текст повинен містити не менше 100 знаків. (Ліміт: 4000) – аналіз розміру тексту.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

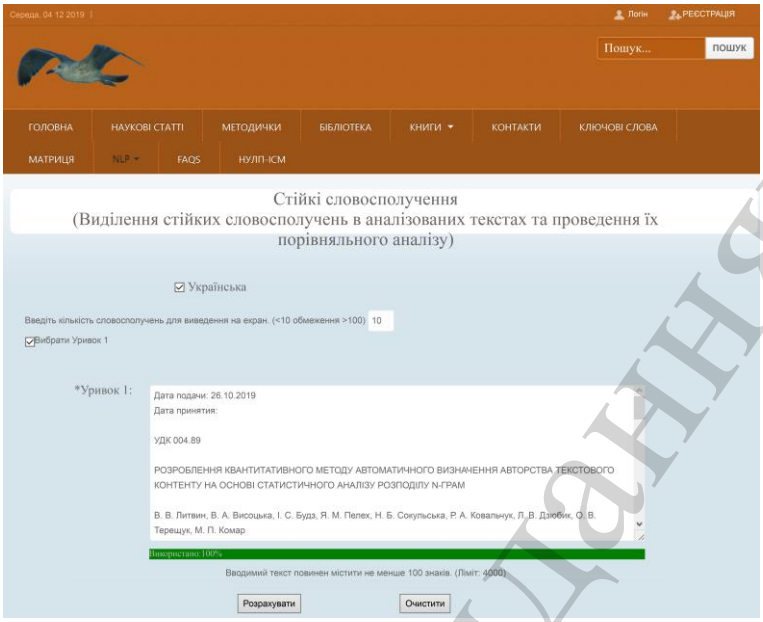


Рис. 9. Приклад застосування аналізу стійких словосполучень

Алгоритм 3. Квантитативний аналіз стійких словосполучень.

Крок. 1. Очищення отриманого контенту від спецсимволів та інш.

Крок. 2. Формуємо список заблокованих слів із бази даних в залежності від вибраної мови контексту.

Крок. 3. Підготовка до формування масивів подвійних словосполучень та всіх слів. На вході масив: ключ – цифри, значення – текст, розбитий по реченням (розділювач крапка). Слова звіряються з базою даних ключових слів та по правилу, описаному в базі даних, приводить дане слово до основи слова, якщо само не є основою слова.

Крок. 4. Визначення стійких словосполучень за методом FREG: отримати абсолютну частоту словосполучень (рис. 10).

Крок. 5. Визначення стійких словосполучень за методом *t*-тест: $P(W1)*P(W2)$ врахування не тільки пар, але і частоти вживання окремих слів (тих, що складають пару).

Крок. 6. Визначення стійких словосполучень за методом LR.

Крок. 7. Визначення стійких словосполучень за методом X2 (табл. 6).

Крок. 8. Результати дослідження виводяться на екран.

Якщо в базі даних відсутнє слово добавляється автоматично. Модератору необхідно для цього слова описати правило приведення слова до основи слова.

При ідентифікації автора тексту передбачається, що текст відображає індивідуальну манеру письма автора, яка дозволяє відрізнити його від інших. Щоб порівнювати тексти між собою необхідно зіставити тексту деяку числову

характеристику, яка була б наближена для текстів одного і того ж автора, і суттєво різнилася б для творів різних авторів. Такою характеристикою може бути щільність функції розподілу (ЩФР) літеросполучень з трьох поспіль символів (3-грам). ЩФР визначається, як сукупність емпіричних частот вживання літер або їх поєднань. При аналізі тексту за допомогою ЩФР не враховують входження розділових знаків, пробілів і цифр. Завдання ідентифікації автора невідомого тексту в термінах ЩФР формулюється так. Дано деякий набір текстів, в якому містяться твори A відомих авторів. Нехай K_a – кількість контенту a-го автор. $N_{i,a}$ – кількість символів в i-му контенті a-го учасника, i=1, ..., K_a. Всі тексти в даному наборі подані у вигляді ЩФР.

№	FREG			t-тест		LR		X2	
1	2			3		4		5	
	Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	logL	Словосполучення	X2
1	ключовий слово	2	0.015873	анализа распределения	1.391766	автор текстовий	6.60e-3	анализа распределения	126.000000
2	основе статистического	2	0.015873	стійкий словосполучення	1.391766	аналізований текст	2.28e-3	стійкий словосполучення	126.000000
3	конкретний автор	2	0.015873	ключовий слово	1.380542	досліджуваний текст	2.28e-3	ключовий слово	83.322581
4	стійкий словосполучення	2	0.015873	основе статистического	1.380542	визначення приналежність	1.00e-3	основе статистического	83.322581
5	автор текстовий	2	0.015873	конкретний автор	1.380542	определения принадлежности	5.54e-4	конкретний автор	83.322581
6	определения принадлежности	2	0.015873	определения принадлежности	1.369318	основе статистического	2.34e-4	определения принадлежности	61.983871
7	аналізований текст	2	0.015873	визначення приналежність	1.358094	ключовий слово	2.34e-4	визначення приналежність	49.180645
8	досліджуваний текст	2	0.015873	аналізований текст	1.335646	конкретний автор	2.34e-4	аналізований текст	34.548387
9	анализа распределения	2	0.015873	досліджуваний текст	1.335646	стійкий словосполучення	3.46e-5	досліджуваний текст	34.548387
10	визначення приналежність	2	0.015873	автор текстовий	1.330034	анализа распределения	3.46e-5	автор текстовий	31.701915

Рис. 10. Приклад результату застосування аналізу стійких словосполучень

ЩФР контенту, обсяг якого дорівнює $N_{i,a}$, задається як множина значень $f_{i,a}(j)=k_j/N_{i,a}$, k_j – кількість вживання N-грами під номером j. Аргумент j=1,...,$a(n,M)$, відповідає номеру літеросполучення (N-грами) при алфавітному впорядкуванні, де M – потужність алфавіту мови написаного тексту, n – порядок N-грами, тобто кількість символів в літеросполученні. $a(n, M)=M^n$ – кількість N-грам в даному алфавіті. Кожен автор ототожнюється з його середньозваженою ЩФР за формулою $F_a(j)=\frac{1}{N_a}\sum_{i=1}^{K_a} f_{i,a}N_{i,a}$. Ці ЩФР є авторськими еталонами. Для порівняння двох текстів, або тексту і авторського еталону, необхідно задати відстань між відповід-

ними функціями розподілу. Як метрики відстані застосовують норму в просторі функцій як доданків. Так, наприклад, відстань $p_{0,a}$ між ЩФР невідомого тексту f_0 і будь-якої авторської ЩФР F_a розраховують як:

$$p_{0,a} = \left\| f_0 - F_a \right\| = \sum_{j=1}^{a(n,M)} \left| f_0(j) - F_a(j) \right|.$$

Відповідно, текст «0» буде належати тому автору, відстань до ЩФР якого буде найменшим. При вирішенні задачі класифікації набір даних не розбивався явно на тестові і тренувальні множини. Середньозважені ЩФР будувалися по всій множині контенту одного автора. Відстань від контенту *i* до конкретного автора *a* обчислювалося як:

$$p_{i,a} = \frac{\left\| f_{i,a} - F_a \right\|}{1 - N_{i,a} / N_a}.$$

Формула дозволяє виключити участь ЩФР контенту *i* в середній ЩФР конкретного автора. На Web-ресурсі для аналізу N-грам є такі поля (рис. 11):

– Вибрати мову тексту – мова тексту для аналізу (дослідження). За замовчуванням «Українська».

– Число грами – кількість знаків у грамі. За замовчуванням 3. Можна міняти на 1, 2, 3, 4.

– Обмеження тексту в знаках.

– Текст – поле, куди копіюється із буфера досліджуваний текст.

– Генерувати – для запуску генерації N-грам.

– Очистити – очищення введених даних.

Алгоритм 4. Лінгвостатистичний аналіз N-грам тексту.

Крок. 1. Очищення досліджуваного тексту (цифри, спецсимволи).

Крок. 2. Вираховуємо кількість слів у тексті.

Крок. 3. Всі слова тексту переводимо в нижній регістр.

Крок. 4. Видаляємо пробіли.

Крок. 5. В залежності від вибраної мови підставляється відповідний алфавіт.

Крок. 6. В залежності від встановленого числа грами запускається відповідна функція, яка розраховує всі можливі варіанти грам і зберігає в масиві.

Крок. 7. Далі запускається функція підрахування кількості входження слів.

Тут же розраховуємо відносну частоту входження та зберігаємо в масиві: порядковий номер грами, сама грама, кількість входжень даної грами, відносна частота входження даної грами.

Крок. 8. Наступна функція формує отриманий в попередній функції масив для експорту в CSV файл. Цей файл зберігається на сервері. Його можна завантажити на комп'ютер користувача (дослідника) по посиланню, доступ до якого буде після формування форми з результатами дослідження.

Крок. 9. Результати дослідження виводяться на екран (тільки ті грами, які знайдено в тексті).

Крок. 10. Відкривається доступ до файлу експорту.

Крок. 11. Виводяться узагальненні результати:

– розмір алфавіту;

– кількість слів у тексті;

– кількість знаків в тексті з пробілами;

– кількість знаків в тексті повністю очищеному;

– всього N-грам;

– всього знайдено N-грам без повторень;

– всього знайдено N-грам з повтореннями.

Таблиця 6

Список за рейтингом частоти появи стійких словосполучень для статті 1, словосполучень: 45. Всього слів: 108

№	FREG			t-тест		LR		X2	
	Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	log L	Словосполучення	X2
1	система електронний	4	0.088889	система електронний	1.822222	інформаційний технологія	5.03e–1	прийняття рішення	45.000000
2	інформаційний система	4	0.088889	електронний контент-комерція	1.578091	інтелектуальний система	2.13e–1	система електронний	45.000000
3	електронний контент-комерція	3	0.066667	розділ науковий	1.319933	інформаційний система	8.36e–2	електронний контент-комерція	32.946429
4	розділ науковий	2	0.044444	інформаційний система	1.222222	портал науковий	5.58e–2	розділ науковий	29.302326
5	портал науковий	1	0.022222	прийняття рішення	0.977778	курс технологія	3.31e–2	курс технологія	21.988636
6	інтелектуальний система	1	0.022222	курс технологія	0.955556	сховище дані	3.31e–2	сховище дані	21.988636
7	прийняття рішення	1	0.022222	сховище дані	0.955556	прийняття рішення	8.27e–3	портал науковий	14.318182
8	курс технологія	1	0.022222	портал науковий	0.933333	розділ науковий	1.89e–3	інформаційний система	5.848550
9	сховище дані	1	0.022222	інтелектуальний система	0.777778	електронний контент-комерція	1.55e–4	інтелектуальний система	3.579545
10	інформаційний технологія	1	0.022222	інформаційний технологія	0.688889	система електронний	1.37e–6	інформаційний технологія	1.890409

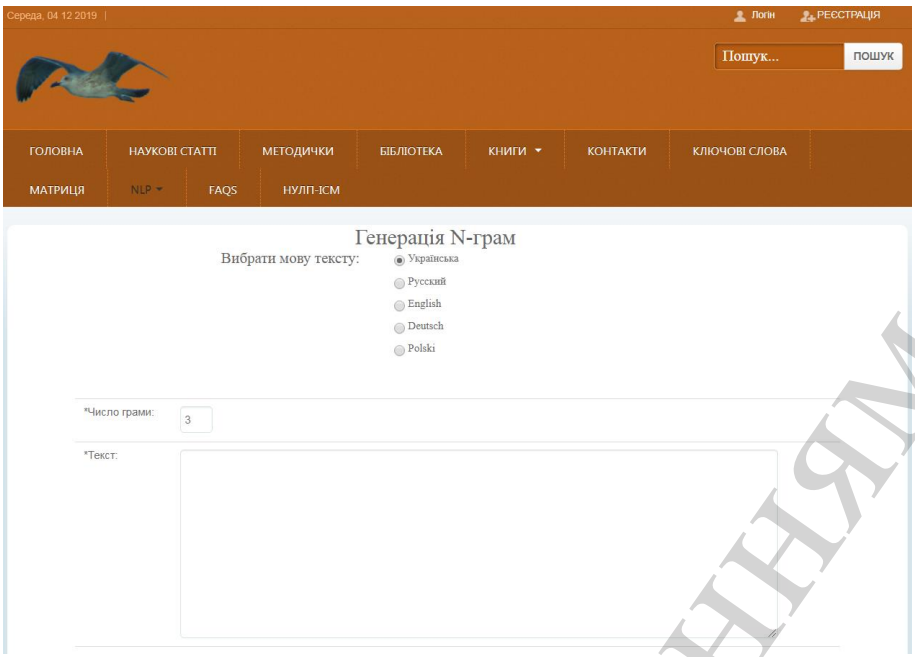


Рис. 11. Приклад застосування аналізу N-грам тексту

6. Результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора тексту

Порівняємо три публікації [1, 74, 77] науково-технічного спрямування між собою на основі лінгвостатистичного аналізу 3-грам. Статті 1, 2 написані одним колективом [1, 74], Стаття 3 написана іншим автором [77] (табл. 7). Мова тексту – українська (літер в алфавіті – 33, тоді всього можливих N-грам 35937).

Таблиця 7

Значення параметрів для аналізованих статей 1–3

Параметри	Стаття 1	Стаття 2	Стаття 3
Всього N-gram	35937	35937	35937
Всього знайдених N-gram (без повторень)	4354	4377	3890
Всього знайдених N-gram (з повторенням)	29494	29862	36383
Всього слів	5475	5358	6060
Всього знаків в неочищеному тексті	39792	39663	47084
Всього знаків в очищеному тексті	29967	32570	37062

Але при порівняні статей будемо враховувати лише ті 3-грами, які зустрілися в тексті одночасно в трьох статтях хоча б один раз. Тому для цього конкретного прикладу всіх 3-грам є 2147. Тобто, для Статті 1 аналізуємо 78,4814 % 3-грам, для Статті 2 – 72,6332 % та для Статті 3 – 84,1271 %. Відповідно різниця вживання відповідних 3-грам між Статтями 1 та 2 є R_{12}=56,5254 %, між Статтями 2 та 3 – R_{23}=69,4271 %, між Статтями 1 та 3 – R_{13}=62.9839 %. Самі ці показники показують, що характеристики статті 1 та 2 більш подібні (R_{23}>R_{12} на 12.9017 %, R_{23} > R_{13} на 6.4432 %, R_{13}> R_{12} на 6.4585 %, тобто R_{23}>R_{13}>R_{12}), ніж характеристики відповідно Статті 1–3 та 2–3. Чим менше R_{ij}, тим більша ступінь, що статті

написані одним і тим же автором. Тоді в випадку Стаття 1 та 2 більш ймовірніше написана одним автором/колективом, ніж Статті 2–3 та Статті 1–3 відповідно. Але проаналізуємо для вживання окремих кластерів 3-грам у відповідних статтях та порівняємо отримані результати. На рис. 12, 13 подано результати аналізу вживання в Статтях 1–3 3-грам, які починаються з літери а (поява в статтях 1–3 в діапазоні 6,1125–6,7087 %). Найчастіше лінії кривих для Статей 1–2 (4,2322 %) та Статей 1–3 (4,197 %) збігаються або наближаються один до одної (середня розбіжність складає 0,02713 % та 0,0269 % відповідно). Але не завжди – є збіг і з Статтею 2–3 (4,6322 %) і є суттєві розбіжності (середня розбіжність складає 0,02969 %). Якщо аналізувати лише такі 3-грами – виходить що всі 3статті написані ймовірніше одним автором. Це пояснюється тим, що ця літера а одною із найчастіше використовуваних для утворення україномовних слів.

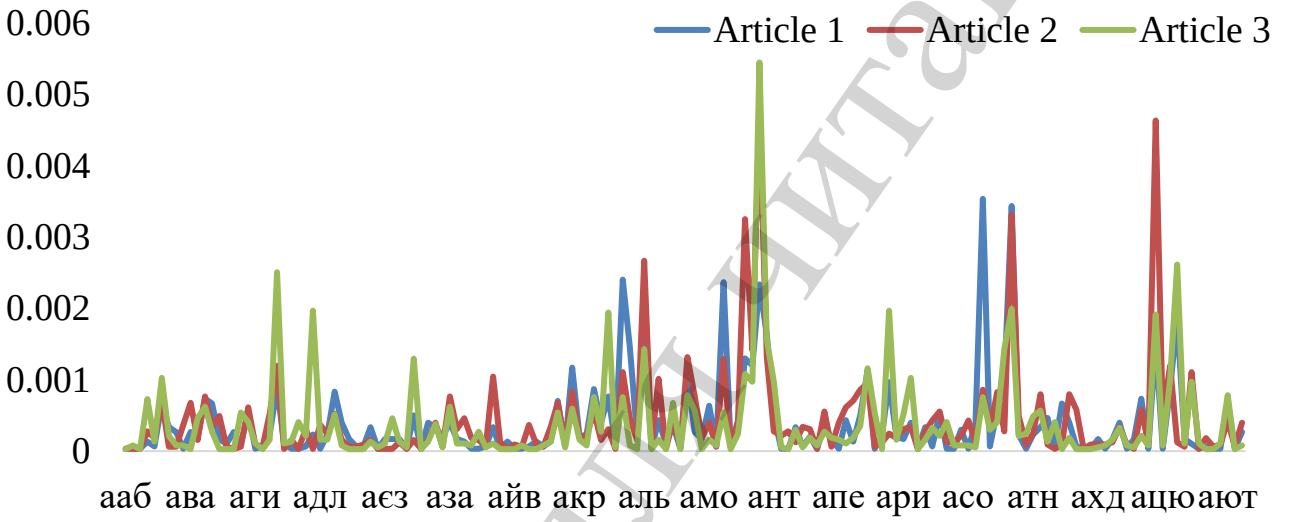


Рис. 12. Графік вживання 3-грам, які починаються з літери а

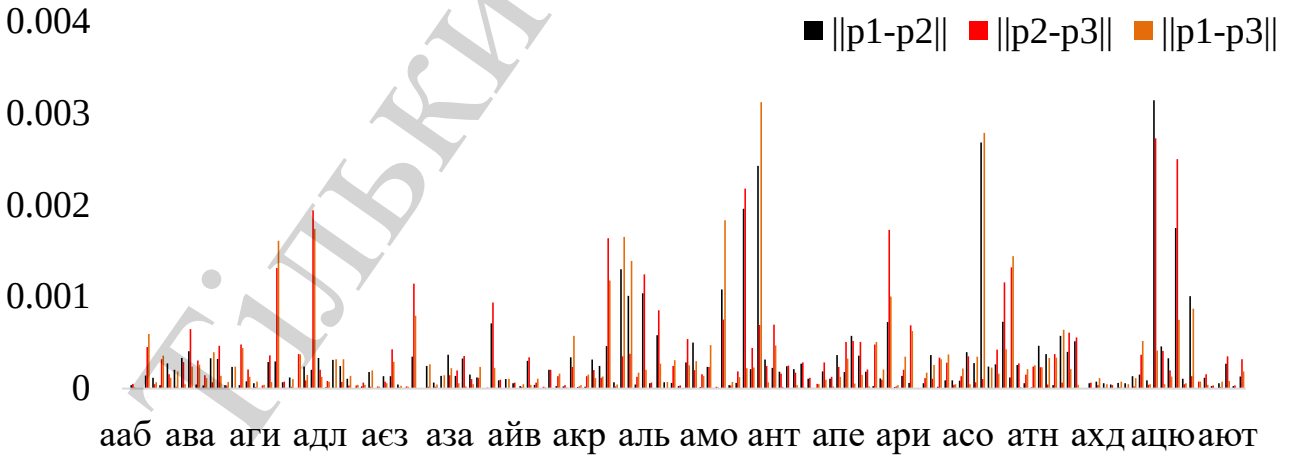


Рис. 13. Графік різниці вживання 3-грам, які починаються з літери а

На рис. 14, 15 подано результати аналізу вживання в Статтях 1–3 3-грам, які починаються з літери б (поява в статтях 1–3 в діапазоні 0,48884–0,77738 %). Найча-

стіше лінії кривих для Статей 1–2 (0,594 %) на відмінну від Статей 1–3 (0,7072 %) та Статей 2–3 (1,1208 %) збігаються або наближаються один до одної. Але траєкторія кривої Статті 1 та Статті 3 найчастіше збігається (ймовірніше статті написані одним автором – середня розбіжність складає 0,01809 %, коли для статей 1–2 – 0,0261 % та статей 2–3 – 0,02866 %). Якщо аналізувати лише такі 3-грами (які рідше зустрічаються) – виходить що всі статті 1–2 написані ймовірніше одним автором, а стаття 3 – іншим. Це пояснюється тим, що ця літера б є рідкою при утворенні україномовних слів. І деякі автори вживають частіше такі слова із-за звички і/або із-за тематики своїх публікації (це потребує додаткових досліджень).

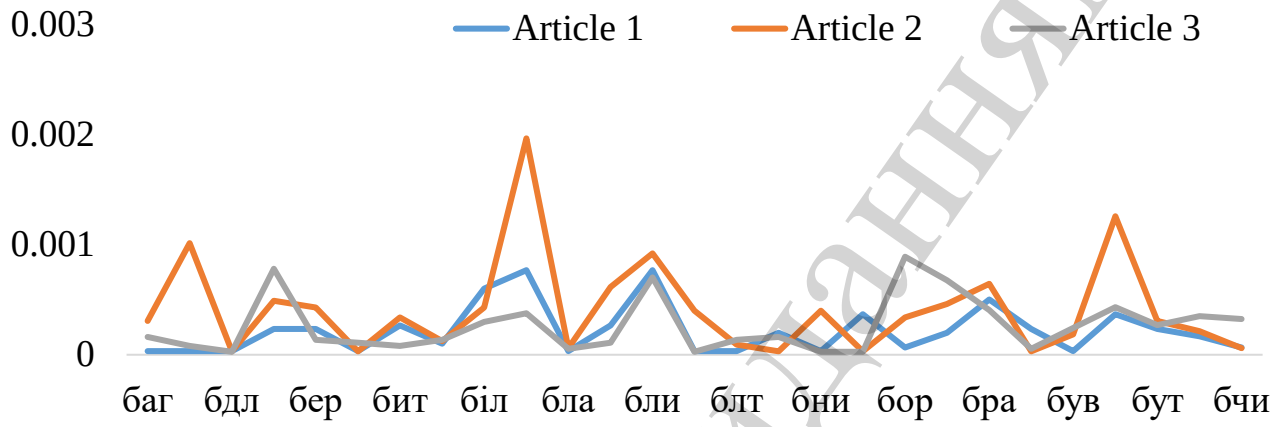


Рис. 14. Графік вживання 3-грам, які починаються з літери б

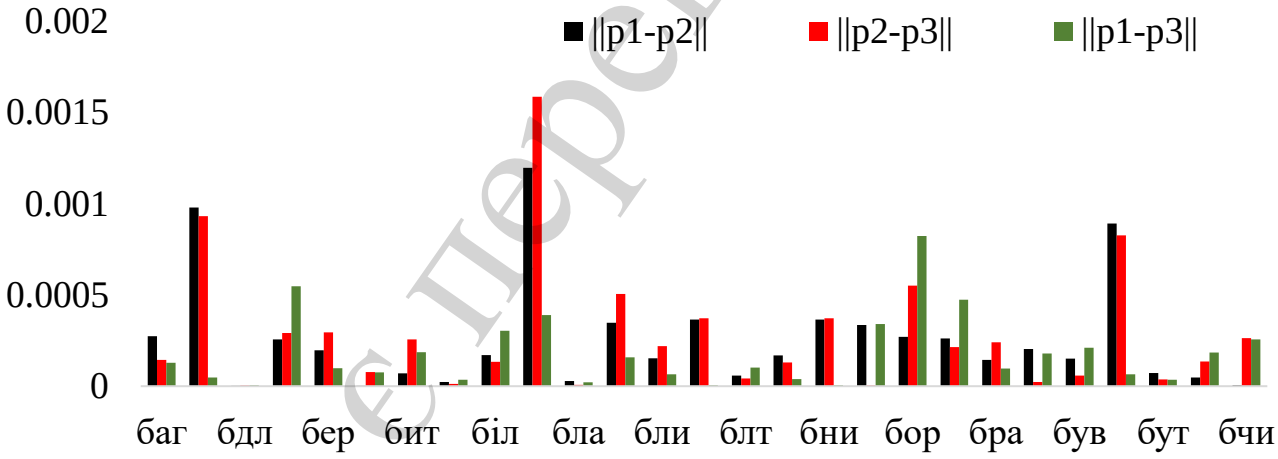


Рис. 15. Графік різниці вживання 3-грам, які починаються з літери б

На рис. 16 подано результати аналізу вживання в Статтях 1–3 3-грам, які починаються з літери в (поява в статтях 1–3 в діапазоні 4,2622–4,5219 %). Найчастіше лінії кривих для Статей 1–2 (3,55581 %), Статей 1–3 (3,6523 %) та Статей 2–3 (4,1064 %) збігаються або наближаються один до одної (середня розбіжність складає 0,03067 %, 0,03149 % та 0,0354 % відповідно). За такими даними – всі три статті ймовірніше написані одним автором.

На рис. 17, 18 подано результати аналізу вживання в Статтях 1–3 3-грам, які починаються з літери г (поява в статтях 1–3 в діапазоні 0,7493–1,4544 %).

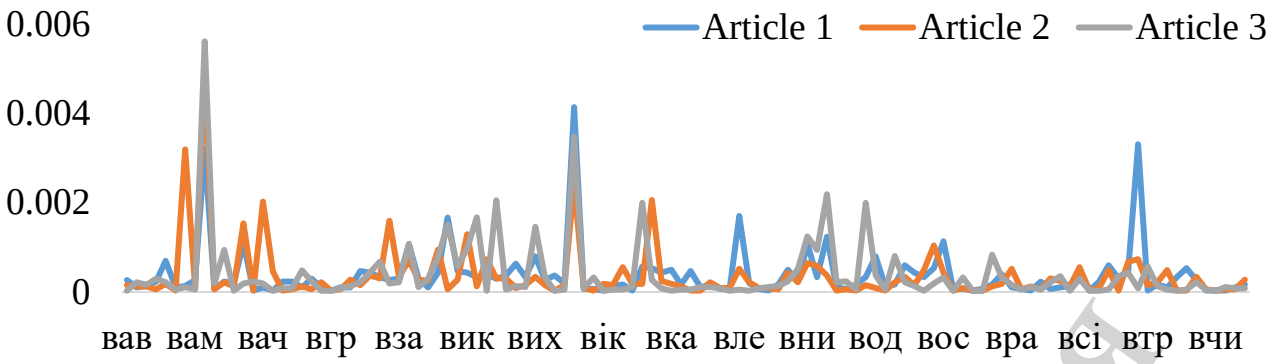


Рис. 16. Графік вживання 3-грам, які починаються з літери в

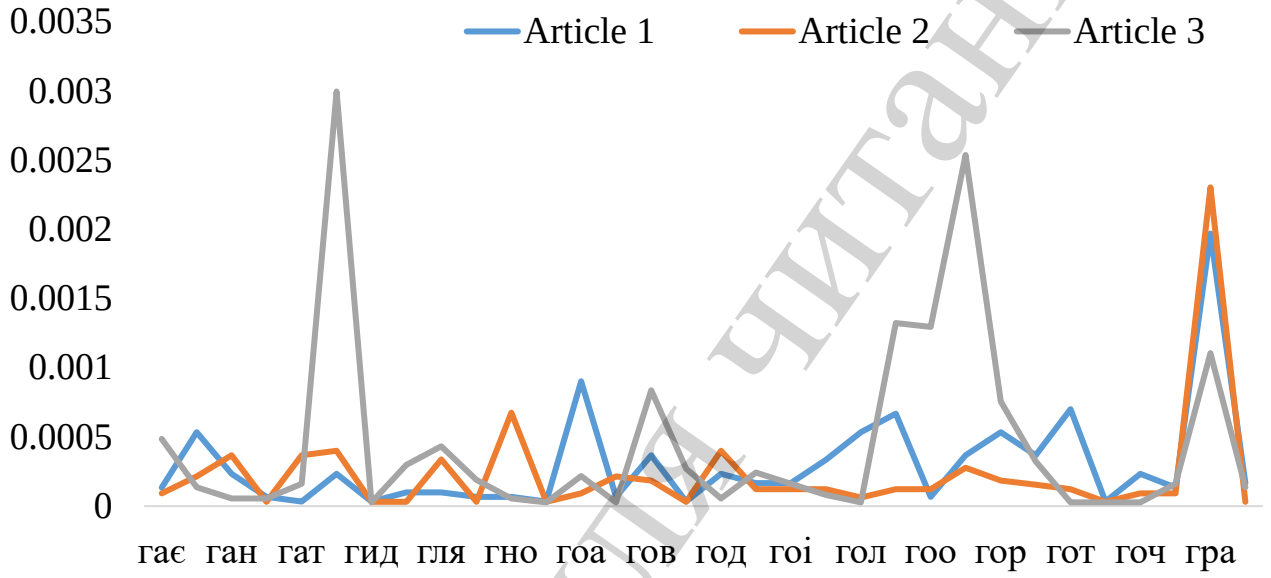


Рис. 17. Графік вживання 3-грам, які починаються з літери г

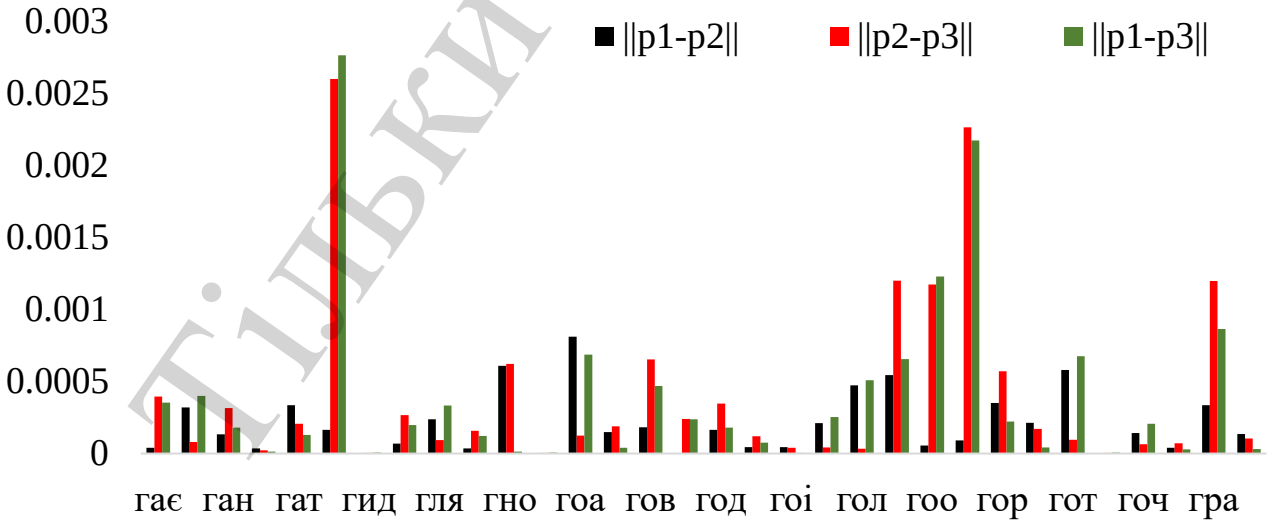


Рис. 18. Графік різниці вживання 3-грам, які починаються з літери г

Найчастіше лінії кривих для Статей 1–2 (0,6551 %) на відмінну від Статей 1–3 (1,309 %) та Статей 2–3 (1,3451 %) збігаються або наближаються один до одної. Але траєкторія кривої Статті 1 та Статті 2 найчастіше збігається (ймовірніше статті написані одним автором, середня розбіжність складає 0,02047 %, коли для статей 2–3 – 0,04203 % та статей 1–3 – 0,04091 %). Якщо аналізувати лише такі 3-грами (які рідше зустрічаються) – виходить що Статті 1–2 написані ймовірніше одним автором, Статті 2–3 та Статті 1–3 – точно написані різними.

7. Обговорення результатів досліджень визначення автора в україномовних текстах на основі технології статистичної лінгвістики

Згідно з даними табл. 8 та рис. 19 частина літер в українській мові найчастіше вживані, інші – набагато рідше. Для найчастіше вживаних літер частота появи 3-грам з такими початковим літерами буде розподіл майже однаковий (пікові значення на графіку рис. 19), а для інших літер – ні.

Рис. 19. Графік розподілу частот появи 1-грами в Статтях 1–3

Тому доцільно досліджувати лише триграми для початкових літер, що рідше зустрічаються в текстах конкретної мови для визначення ступеня належності тексту відповідному автору (наприклад, рис. 120, 21). Так, для 3-грами з літери є (поява в статтях 1–3 в діапазоні 0,2517–0,707 %) найчастіше лінії кривих для Статей 1–2 (0,2508 %) на відмінну від Статей 1–3 (0,6077 %) та Статей 2–3 (0,5443 %) збігаються або наближаються один до одної. Але траєкторія кривої Статті 1 та Статті 2 найчастіше збігається (ймовірніше статті написані одним автором – середня розбіжність складає 0,0114 %, коли для Статей 2–3 – 0,02478 % та Статей 1–3 – 0,02762 % це значення більше в 2 рази).

Але це спрацьовує не завжди. Так для 3-грами з літери ж (поява в статтях 1–3 в діапазоні 0,3408–0,4738 %) всі лінії кривих для Статей 1–2 (0,25 %), Статей 1–3 (0,2126 %) та Статей 2–3 (0,2302 %) збігаються або наближаються один до одної. Середня розбіжність для Статей 1–2 складає 0,01786 %, коли для Статей 2–3 – 0,01644 % та Статей 1–3 – 0,01519 %. Ніби всі статті написані одним

автором. Хоча траєкторія кривих на рис. 22 та стовпці діаграми на рис. 23 показують, що швидше Статті 1–2 написані одним автором, а Стаття 3 – іншим.

Таблиця 8
Розподіл частот появи 1-грами в Статтях 1–3

№	1-грама	Стаття 1		Стаття 2		Стаття 3	
		Кількість	ВЧ	Кількість	ВЧ	Кількість	ВЧ
1	а	2255	0.075252	2698	0.082837	2491	0.066685
2	б	284	0.009477	569	0.017470	428	0.011458
3	в	1654	0.055196	1590	0.048818	1915	0.051265
4	г	408	0.013615	373	0.011452	651	0.017427
5	д	859	0.028666	939	0.028830	1319	0.035310
6	е	1404	0.046853	1453	0.044612	2090	0.055950
7	є	188	0.006274	165	0.005066	347	0.009289
8	ж	246	0.008209	210	0.006448	176	0.004712
9	з	623	0.020790	644	0.019773	946	0.025325
10	и	1732	0.057799	1852	0.056862	2036	0.054504
11	і	1789	0.059701	1967	0.060393	2250	0.060233
12	ї	174	0.005807	217	0.006663	270	0.007228
13	й	239	0.007976	260	0.007983	265	0.007094
14	к	1279	0.042682	1110	0.034080	1453	0.038897
15	л	1116	0.037242	927	0.028462	906	0.024254
16	м	808	0.026964	976	0.029966	1399	0.037451
17	н	2471	0.082460	2370	0.072766	2888	0.077312
18	о	2824	0.094240	2472	0.075898	3870	0.103601
19	п	647	0.021591	825	0.025330	1138	0.030464
20	р	1335	0.044550	1722	0.052871	1893	0.050676
21	с	1549	0.051692	1327	0.040743	1384	0.037050
22	т	2102	0.070146	1956	0.060055	2141	0.057315
23	у	987	0.032937	960	0.029475	1195	0.031990
24	ф	179	0.005973	209	0.006417	137	0.003668
25	х	355	0.011847	384	0.011790	482	0.012903
26	ц	224	0.007475	334	0.010255	299	0.008004
27	ч	459	0.015317	289	0.008873	574	0.015366
28	ш	117	0.003904	169	0.005189	281	0.007522
29	щ	95	0.003170	52	0.001597	128	0.003427
30	ь	498	0.016619	418	0.012834	613	0.016410
31	ю	156	0.005206	277	0.008505	289	0.007737
32	я	647	0.021591	681	0.020909	864	0.023129

Перевіримо ще раз. Для 3-грами з літери з (поява в статтях 1–3 в діапазоні 1,3108–1,973 %) лінії кривих для Статей 1–2 (1,1879 %), Статей 1–3 (1,3259 %) та Статей 2–3 (1,25 %) збігаються або наближаються один до одної. Середня розбіжність для Статей 1–2 складає 0,02121 %, коли для Статей 2–3 – 0,02232 %

та Статей 1–3 – 0,02368 %. Ніби всі статті написані одним автором. Хоча траєкторія кривих на рис. 24 показує, що швидше Статті 1 та 2 написані одним автором, а Стаття 3 – іншим.

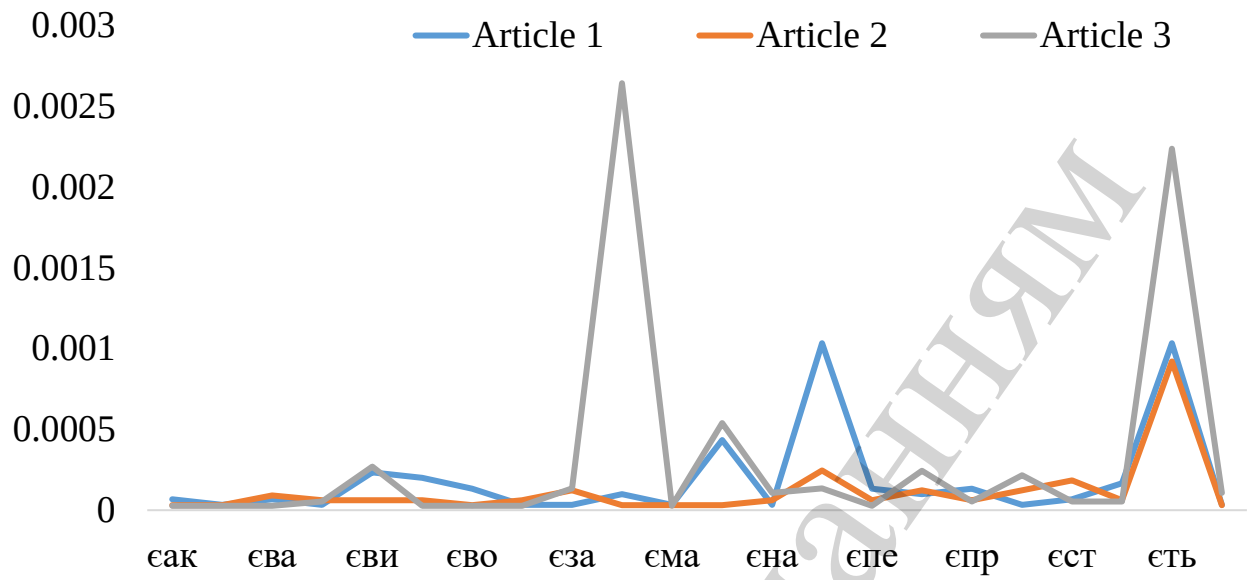


Рис. 20. Графік вживання 3-грам, які починаються з літери є

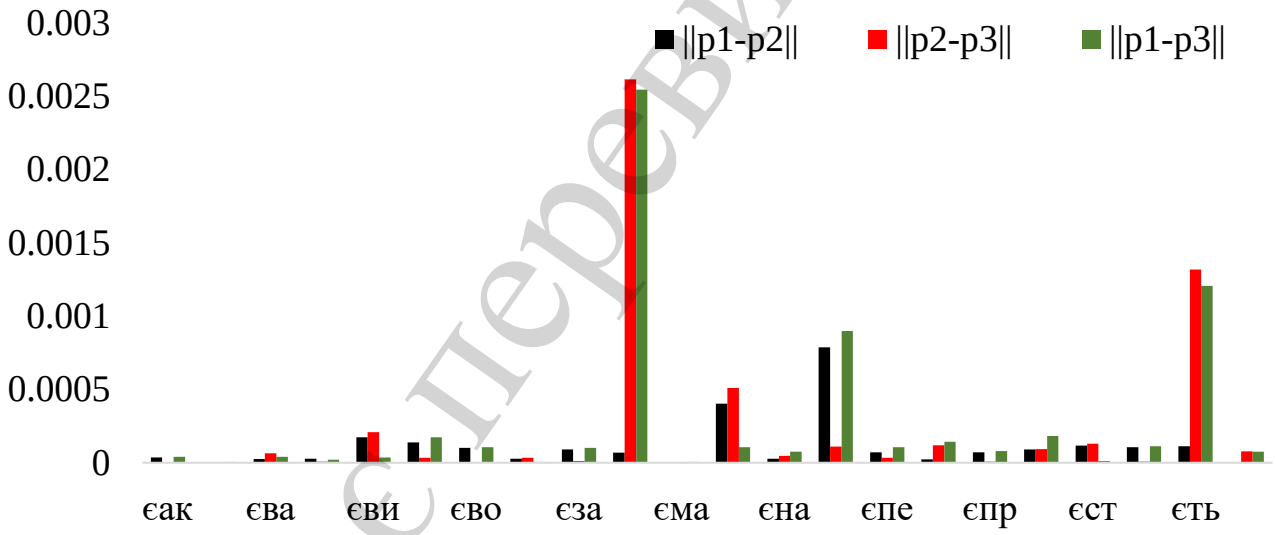


Рис. 21. Графік різниці вживання 3-грам, які починаються з літери є

Для 3-грами з літери й (поява в статтях 1–3 в діапазоні 0,301–0,4319 %) лінії кривих для Статей 1–2 (0,3352 %), Статей 1–3 (0,3483 %) та Статей 2–3 (0,3469 %) збігаються або наближаються один до одної. Середня розбіжність для Статей 1–2 складає 0,01457 %, коли для Статей 2–3 – 0,01508 % та Статей 1–3 – 0,01514 %. Ніби всі Статті написані одним автором. Хоча траєкторія кривих на рис. 25 показує, що скоріше всього Статті 1–2 написані одним автором, а Стаття 3 – іншим.

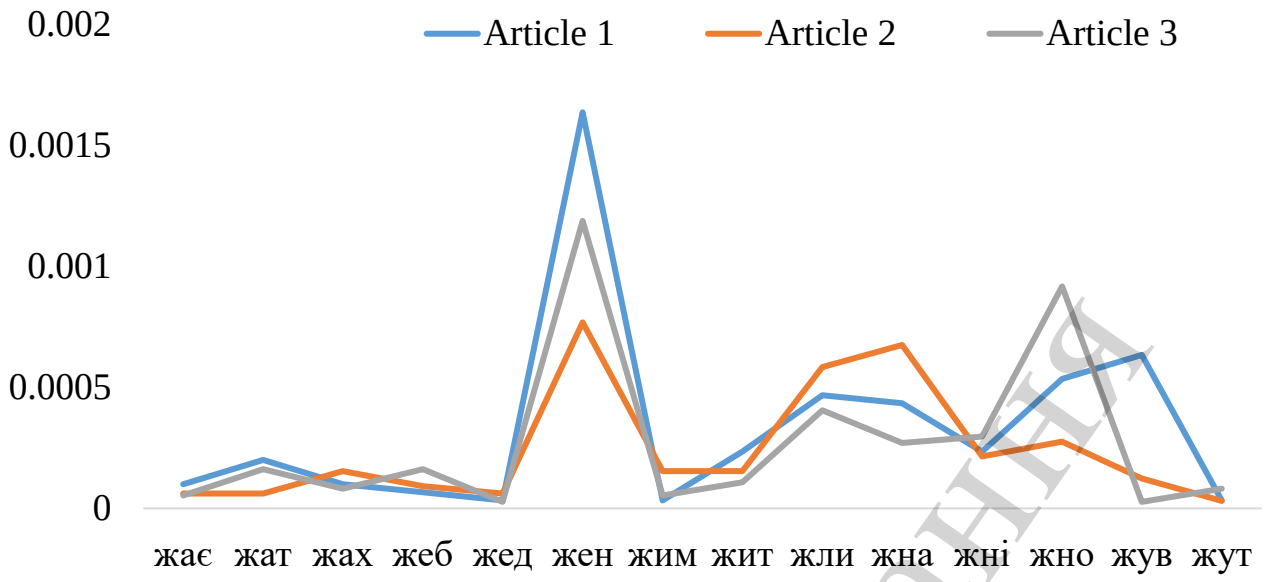


Рис. 22. Графік вживання 3-грам, які починаються з літери ж

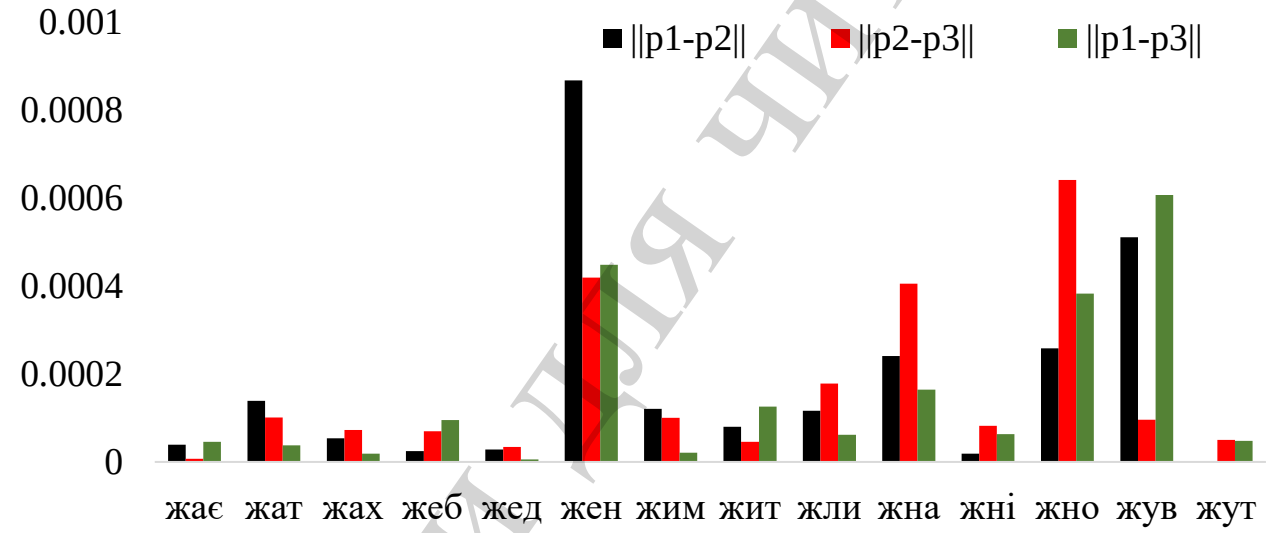


Рис. 23. Графік різниці вживання 3-грам, які починаються з літери ж

Для 3-грами з літери м (поява в статтях 1–3 в діапазоні 2,1681–3,1225 %) лінії кривих для Статей 1–2 (1,7619 %) та Статей 1–3 (1,8193 %) на відмінну від Статей 2–3 (2,6606 %) збігаються або наближаються один до одної. Середня розбіжність для Статей 1–2 складає 0,01936 %, коли для Статей 2–3 – 0,02936 % та Статей 1–3 – 0,02 %. Ніби всі Статті написані одним автором. Хоча траєкторія кривих на рис. 26 показує, що швидше Статті 1 та 2 написані одним автором, а Стаття 3 – іншим. Отже не лише кількість появи триграм з певною літерою на початку впливає на коректність результату, але і частота появи таких 3-грам.

Для 3-грами з літери п (поява в статтях 1–3 в діапазоні 1,8583–2,8092 %) всі лінії кривих для Статей 1–2 (1,6619 %), на відмінну від Статей 1–3 (2,1261 %) та Статей 2–3 (2,5456 %) збігаються або наближаються один до одної (рис. 27). Середня розбіжність для Статей 1–2 складає 0,04261 %, коли для

статей 2–3 – 0,06527 % та статей 1–3 – 0,05452 %. Статті 1–2 написанні одним автором, а Стаття 3 – іншим.

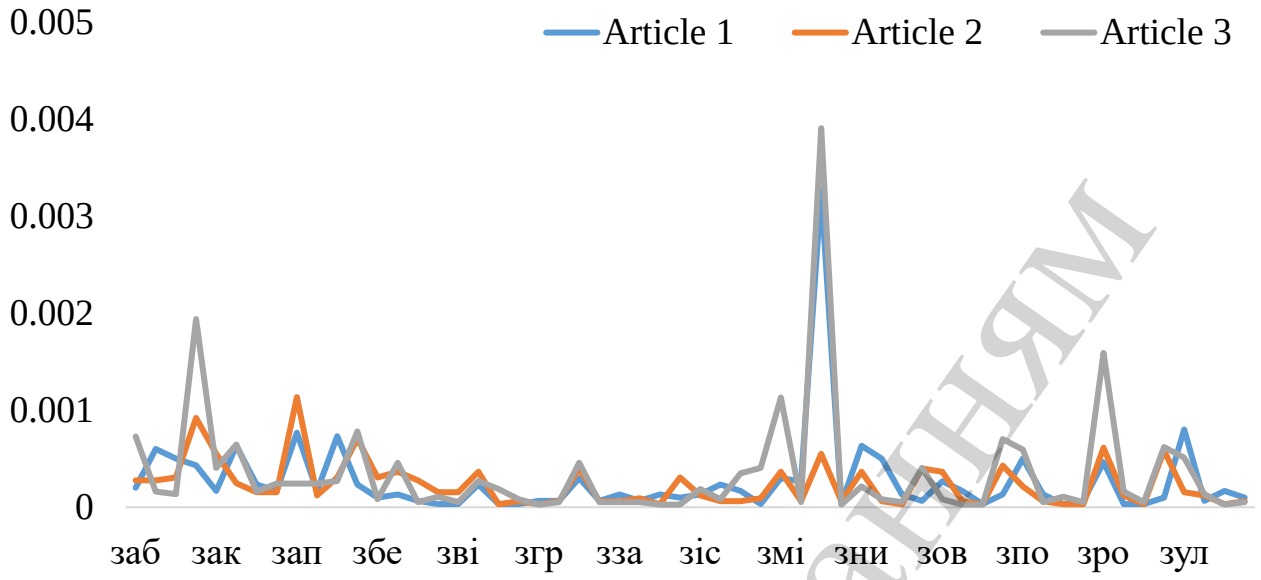


Рис. 24. Графік вживання 3-грам, які починаються з літери з

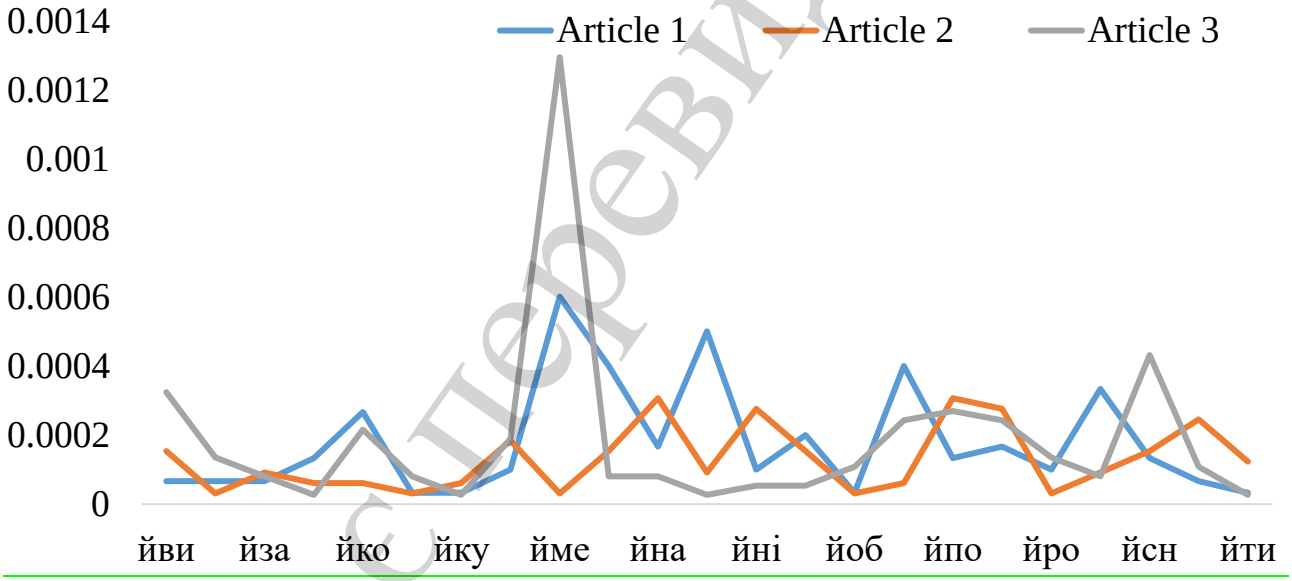


Рис. 25. Графік вживання 3-грам, які починаються з літери й

Для 3-грами з літери р (поява в статтях 1–3 в діапазоні 3,69–4,3802 %) всі лінії кривих для Статей 1–2 (3,1902 %), Статей 1–3 (3,4834 %) та Статей 2–3 (4,3566 %) збігаються або наближаються один до одної (рис. 28). Середня розбіжність для Статей 1–2 складає 0,03323 %, коли для статей 2–3 – 0,04538 % та статей 1–3 – 0,03629 %. Ніби всі статті написані одним автором.

Для 3-грами з літери у (поява в статтях 1–3 в діапазоні 2,1927–2,7261 %) всі лінії кривих для Статей 1–2 (1,7905 %), на відмінну від Статей 1–3 (1,9443 %) та Статей 2–3 (1,9852 %) збігаються або наближаються один до одної (рис. 29). Середня розбіжність для Статей 1–2 складає 0,02184 %, коли для

статей 2–3 – 0,02421 % та статей 1–3 – 0,02371 %. Ніби всі статті написані одним автором. Як триграма з конкретної літери вживається в тексті понад 1 %, тоді середня розбіжність при порівнянні декількох статей, незалежно від авторства, будуть майже однакові. Тому необхідно лише враховувати при аналізі триграми, які мають відсоток появи менший за 1.

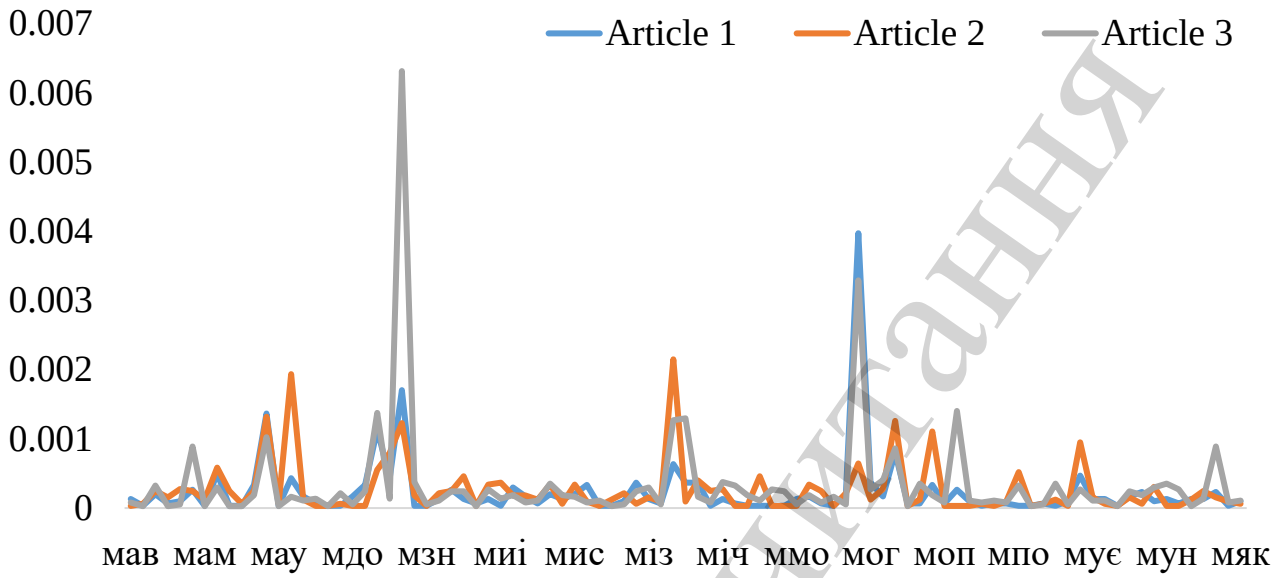


Рис. 26. Графік вживання 3-грам, які починаються з літери м

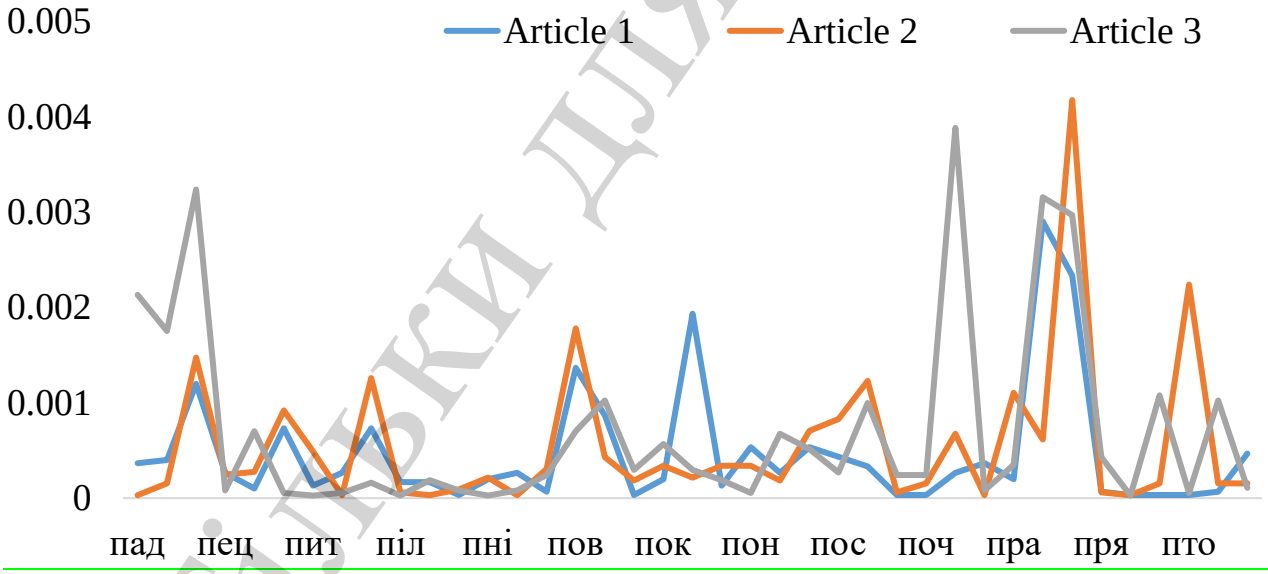


Рис. 27. Графік вживання 3-грам, які починаються з літери п

Для 3-грами з літери ф (поява в статтях 1–3 в діапазоні 0,3069–0,4759 %) всі лінії кривих для Статей 1–2 (0,2762 %) та Статей 1–3 (0,299 %) на відмінну Статей 2–3 (0,495 %) збігаються або наближаються один до одної (рис. 30 Середня розбіжність для Статей 1–2 складає 0,03453 %, коли для статей 2–3 – 0,06188 % та статей 1–3 – 0,03738 %. Ніби статті 1–2 написані одним автором, аналогічно статті 1–3 написані одним автором, а статті 1–3 написані різними авторами. Але траєкторія кри-

вих збігається для Статей 1–2. Хоч частота появи триграми з літерою ф на початку менша 1 % в кожній статті (це дозволило розбити множину потенційних авторів на дві підмножини), але частота появи кожної триграми на літеру ф досить мала (8 триграм із 333 можливих). У порівнянні з літерою а – 156 триграм із 333 можливих. Найкращі результати дають триграми з конкретною літерою на початку, коли їх кількість в межах (30,90). Ці числа є наближеними. Для уточнення їх значень при дослідженні україномовних науково-технічних текстів необхідно провести додаткове дослідження на досить великому обсязі текстів (понад 1000) серед великої кількості авторів (понад 100) та мати точно еталонні одноосібні авторські тексти цих авторів (із підтвердженням їх авторства). Що зробити майже неможливо, у зв'язку з тим, що більшість науково-технічної літератури пишеться у співавторстві з іншими авторами. Це накладає суб'єктивні характеристики на аналізований текст.

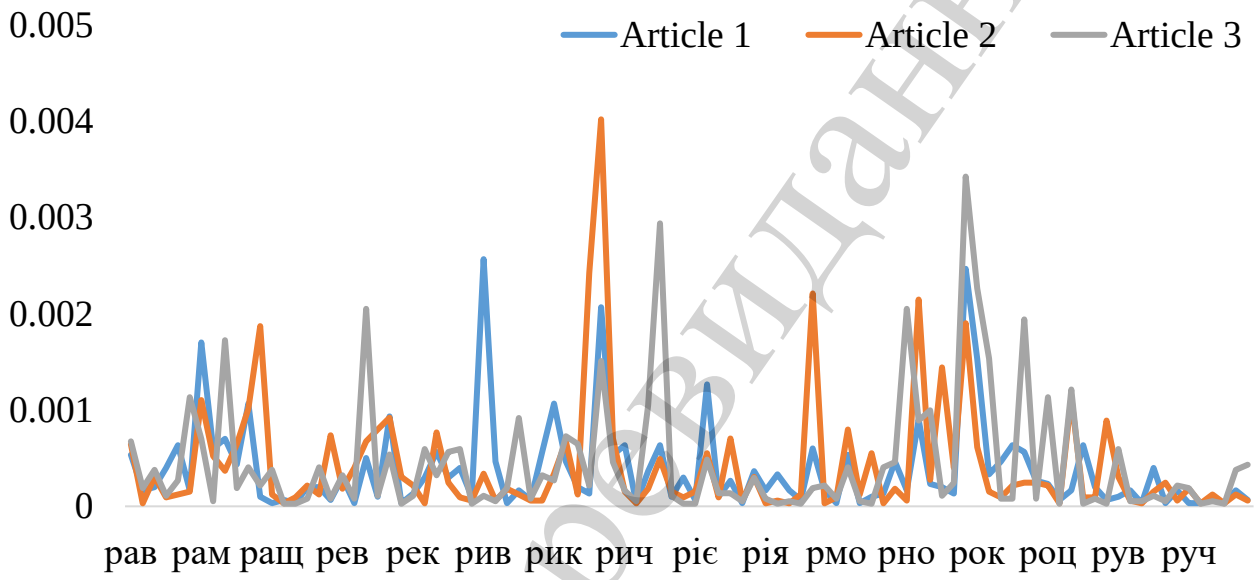


Рис. 28. Графік вживання 3-грам, які починаються з літери р

Перевіримо на практиці. Для 3-грами з літери х (поява в статтях 1–3 в діапазоні 0,5732–0,9339 %, всього 37 триграми) всі лінії кривих для Статей 1–2 (0,5083 %), на відмінну від Статей 1–3 (0,7957 %) та Статей 2–3 (0,7426 %) збігаються або наближаються один до одної (рис. 31). Середня розбіжність для Статей 1–2 складає 0,01374 %, коли для статей 2–3 – 0,02007 % та статей 1–3 – 0,02151 %. Статті 1–2 написані одним автором, а Стаття 3 – іншим.

Для 3-грами з літери ц (поява в статтях 1–3 в діапазоні 0,5906–0,829 %, всього 24 триграми) всі лінії кривих для Статей 1–2 (0,568 %), Статей 1–3 (0,4748 %) та Статей 2–3 (0,4416 %) збігаються або наближаються один до одної (рис. 32). Середня розбіжність для Статей 1–2 складає 0,02367 %, коли для статей 2–3 – 0,0184 % та статей 1–3 – 0,01978 %. Ніби всі статті написані одним автором.

Для 3-грами з літери ч (поява в статтях 1–3 в діапазоні 0,5128–1,3244 %) всі лінії кривих для Статей 1–2 (1,0044 %), Статей 1–3 (0,6924 %) та Статей 2–3 (0,9368 %) збігаються або наближаються один до одної (рис 33). Середня роз-

біжність для Статей 1–2 складає 0,04367 %, коли для статей 2–3 – 0,04073 % та статей 1–3 – 0,0301 %. Ніби всі статті написані одним автором.

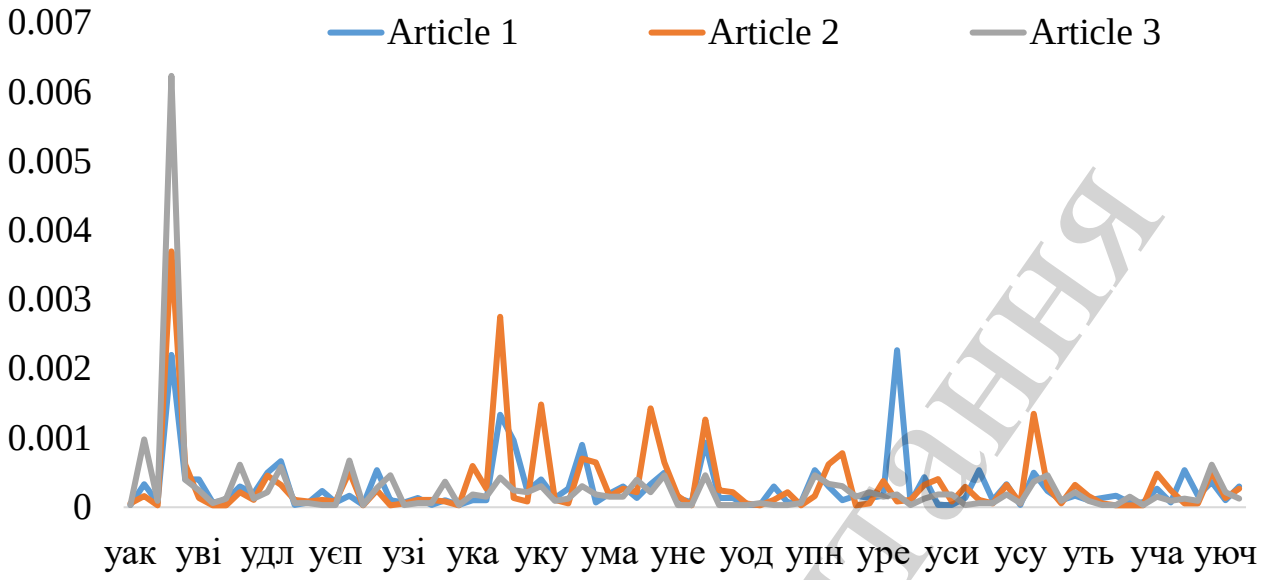


Рис. 29. Графік вживання 3-грам, які починаються з літери у

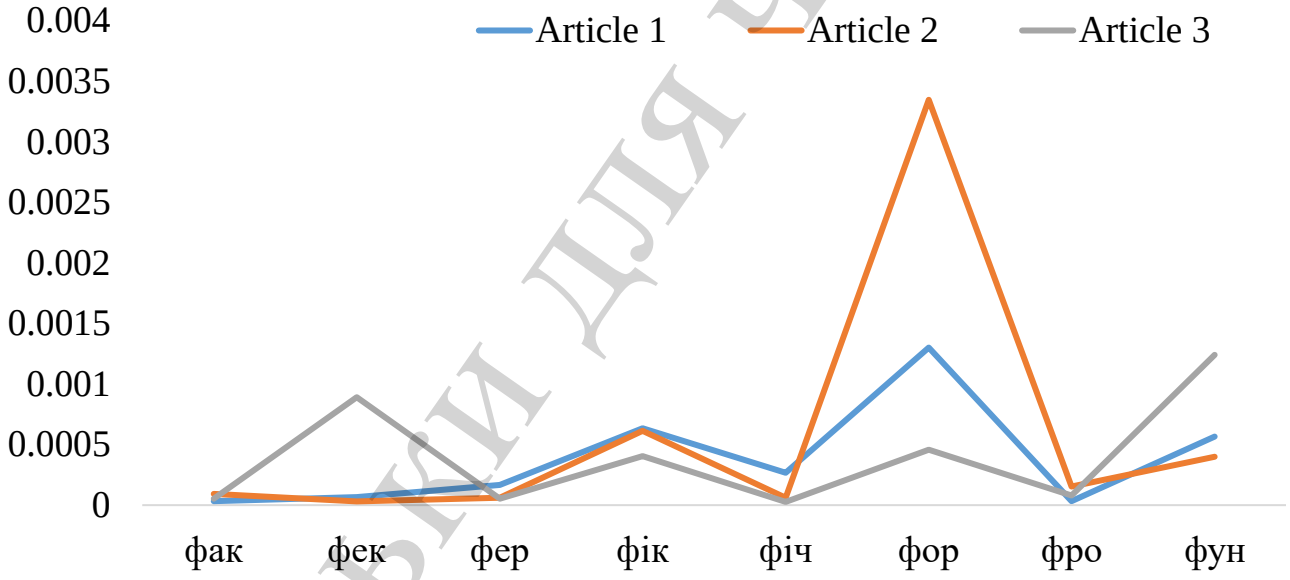


Рис. 30. Графік вживання 3-грам, які починаються з літери ф

Для 3-грами з літери ь (поява в статтях 1–3 в діапазоні 0,9981–1,2848 %, всього 30 триграм) всі лінії кривих для Статей 1–2 (0,6593 %), Статей 1–3 (0,7326 %) та Статей 2–3 (0,7983 %) збігаються або наближаються один до одної (рис. 34). Середня розбіжність для Статей 1–2 складає 0,01691 %, коли для статей 2–3 – 0,02047 % та статей 1–3 – 0,01878 %. Статті 1–2 написанні одним автором, а Стаття 3 – іншим. Але самі значення є межовими із того, що поява триграми з літери ь в діапазоні (0,9;1).

Для 3-грами з літер ш-щ (поява в статтях 1–3 в діапазоні 0,357–0,8258 %, всього триграм 22) всі лінії кривих для Статей 1–2 (0,2625 %), на відмінну від

Статей 1–3 (0,667 %) та Статей 2–3 (0,7209 %) збігаються або наближаються один до одної (рис. 35). Середня розбіжність для Статей 1–2 складає 0,01193 %, коли для статей 2–3 – 0,03277 % та статей 1–3 – 0,03032 %. Точно статті 1–2 написані одним автором, а Стаття 3 – іншим. Це пояснюється тим, що порівнюємо дві різні множини триграм з двома різними початковими літерами.

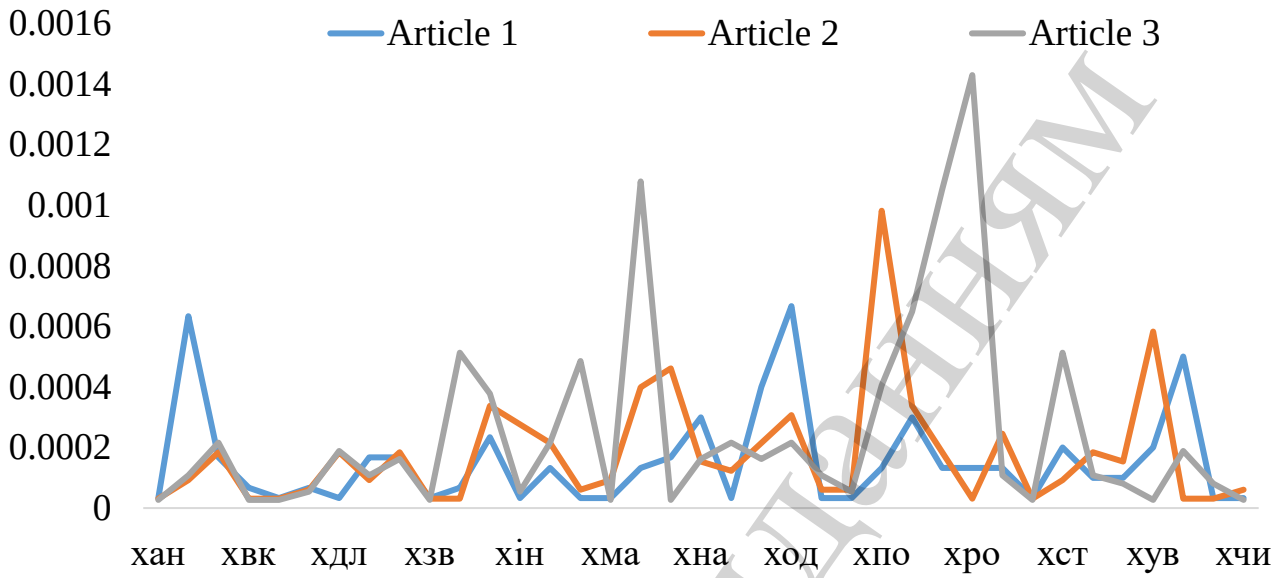


Рис. 31. Графік вживання 3-грам, які починаються з літери х

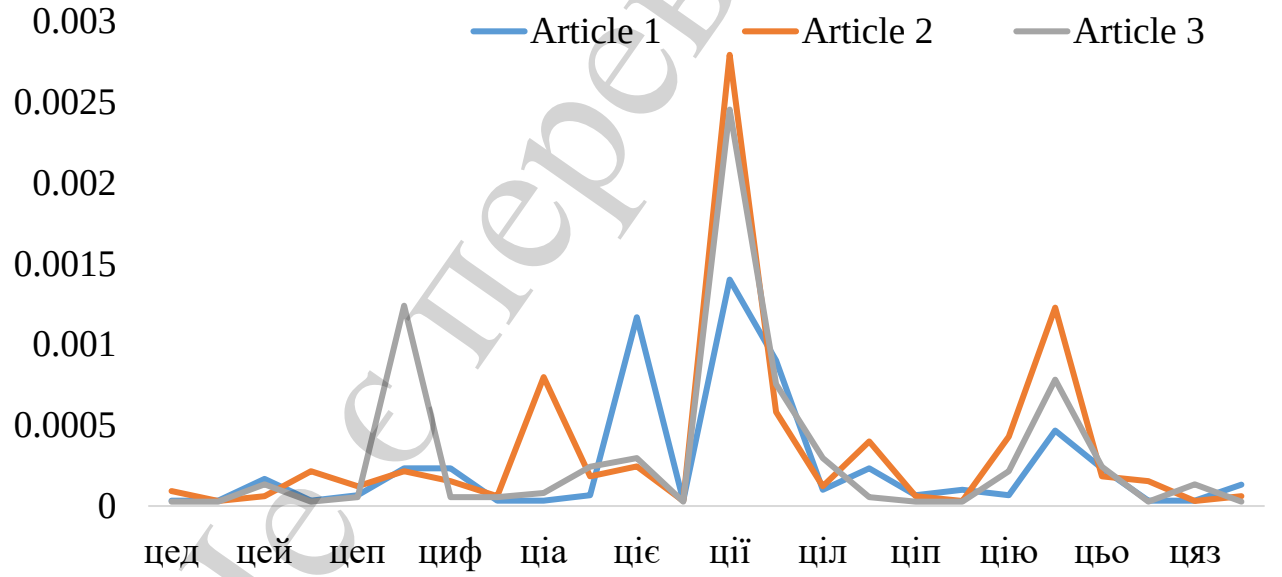


Рис. 32. Графік вживання 3-грам, які починаються з літери ц

Для 3-грами з літери ю (поява в статтях 1–3 в діапазоні 0,2768–0,4939 %) всі лінії кривих для Статей 1–2 (0,1558 %), на відмінну від Статей 1–3 (0,2673 %) та Статей 2–3 (0,2005 %) збігаються або наближаються один до одної (рис. 36). Середня розбіжність для Статей 1–2 складає 0,0097375 %, коли для статей 2–3 – 0,01878 % та статей 1–3 – 0,01671 %. %. Точно статті 1–2 на-

писані одним автором, а Стаття 3 – іншим. Це пояснюється тим, що частота появи таких триграм значно менша 1 %, а точніше менша навіть за 0,5 %.

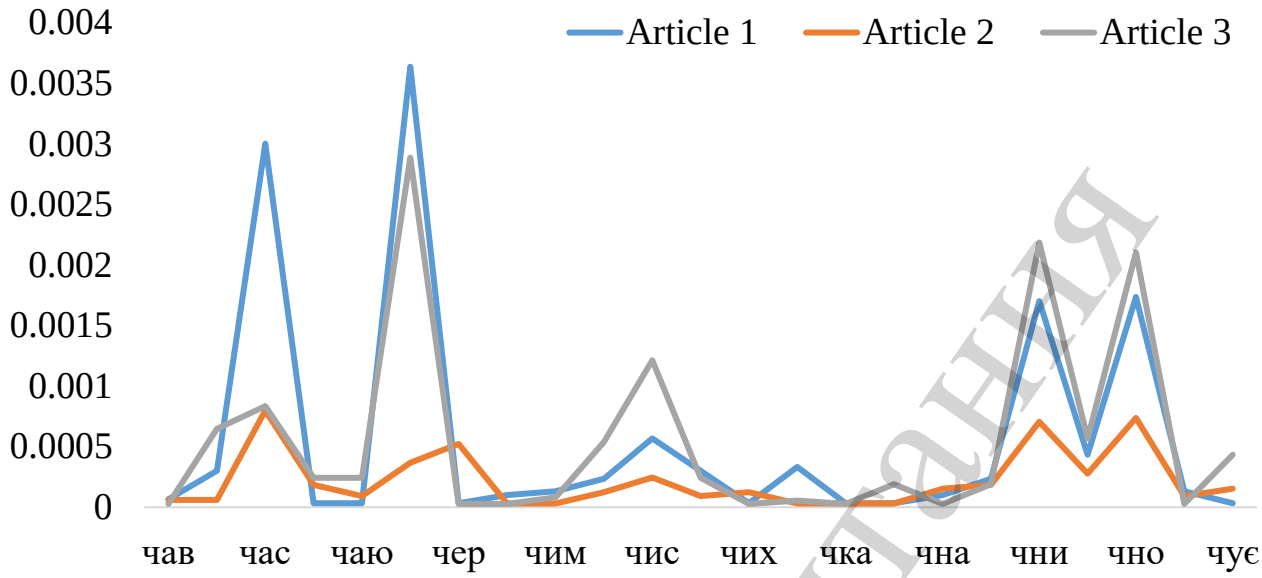


Рис. 33. Графік вживання 3-грам, які починаються з літери ч

Рис. 34. Графік вживання 3-грам, які починаються з літери ь

Для 3-грами з літери я (поява в статтях 1–3 в діапазоні 1,4442–1,5541 %, всього 72 триграми) всі лінії кривих для Статей 1–2 (0,9522 %), Статей 1–3 (0,9361 %) та Статей 2–3 (1,0555 %) збігаються або наближаються один до одної (рис. 37). Середня розбіжність для Статей 1–2 складає 0,013225 %, коли для статей 2–3 – 0,01466 % та статей 1–3 – 0,013 %. Ніби всі статті написані одним автором. Це пояснюється тим, що частота появи таких триграм значно більша за 1 %.

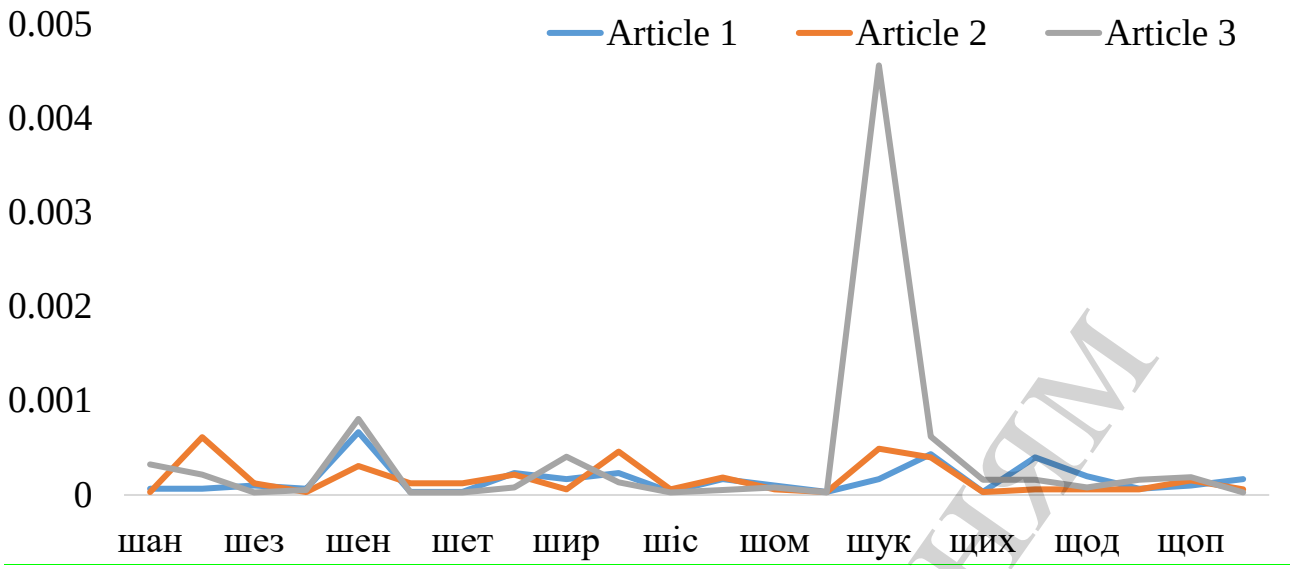
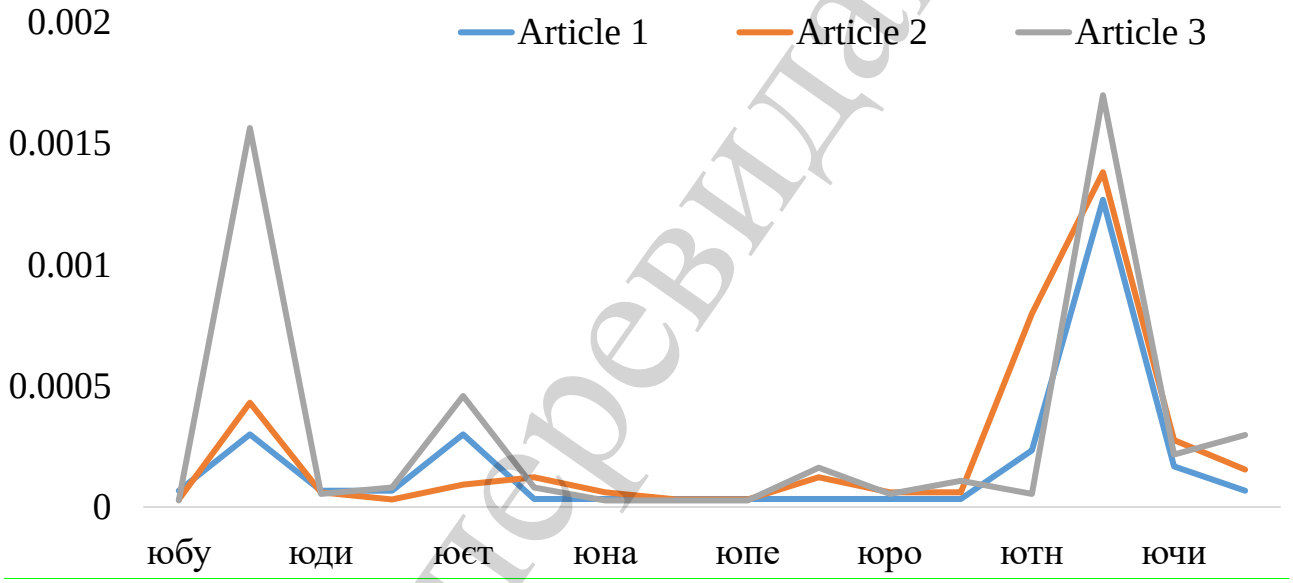


Рис. 35. Графік вживання 3-грам, які починаються з літер ш та щ



с. 36. Графік вживання 3-грам, які починаються з літери ю

Згідно цих графіків випливає, що Стаття 1 та Стаття 2 ймовірніше були написані одним автором, хоча Стаття 1 та Стаття також могли бути написані одним автором (але це не є істиною). А ось статті 2–3 точно були написані різними авторами. Застосування лінгвостатистичного аналізу 3-грам до множини статей дозволить сформувати підмножину подібних за лінгвістичними характеристиками публікацій. Накладання на цю підмножину додаткових умов у вигляді проведення статистичних та квантитативних аналізів (множини ключових слів, стійких словосполучень, стилеметричного, лігвометричного тощо) дозволить значно скоротити цю підмножину, уточнивши список ймовірніших авторських робіт. Так, аналіз змісту та частоти появи лише службових слів відокремить статті 1та 3 в різні підмножини, статті 1 та 2 залишить в одній.

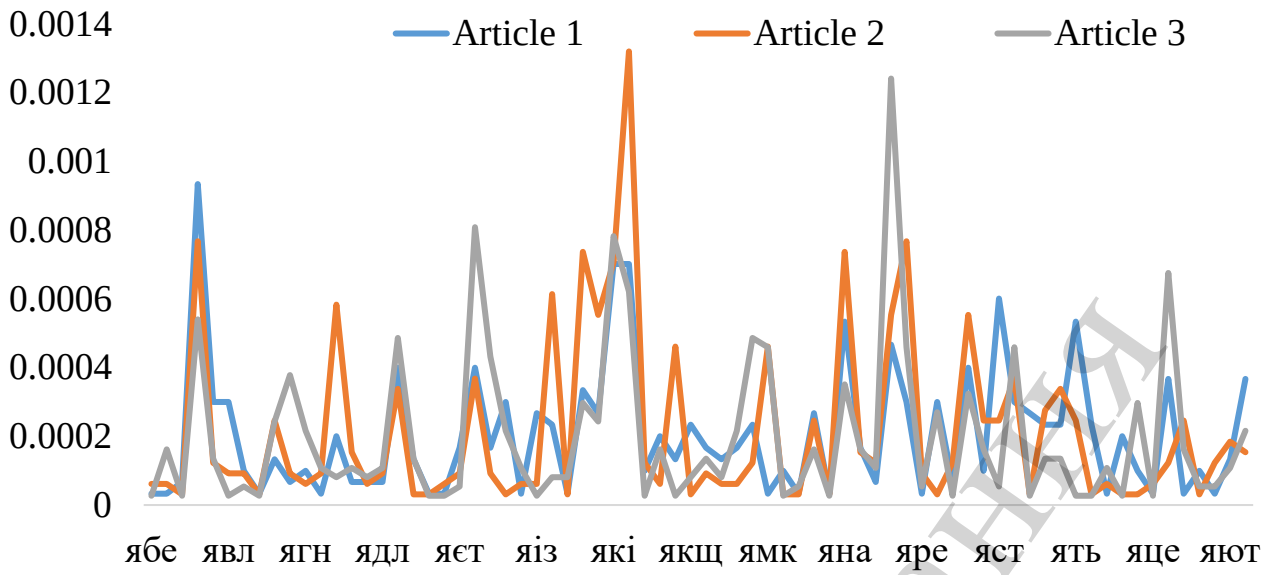


Рис. 37. Графік вживання 3-грам, які починаються з літери я

Тому порівняємо частоти появи всіх триграм, які починаються конкретної літери (рис. 38, 39).

Рис. 38. Графік вживання 3-грам, які починаються з конкретної літери

Дане дослідження не передбачає вирішення завдання ідентифікації автора в повному обсязі з тієї причини, що відмінність авторських рис носить суб'єктивний характер і залежить від обмежень, що накладаються на творчий процес автора. Однак в результаті система, яка реалізує подібні методи, здатна давати рекомендації про ступінь приналежності тексту конкретному автору. Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення стилю автора з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

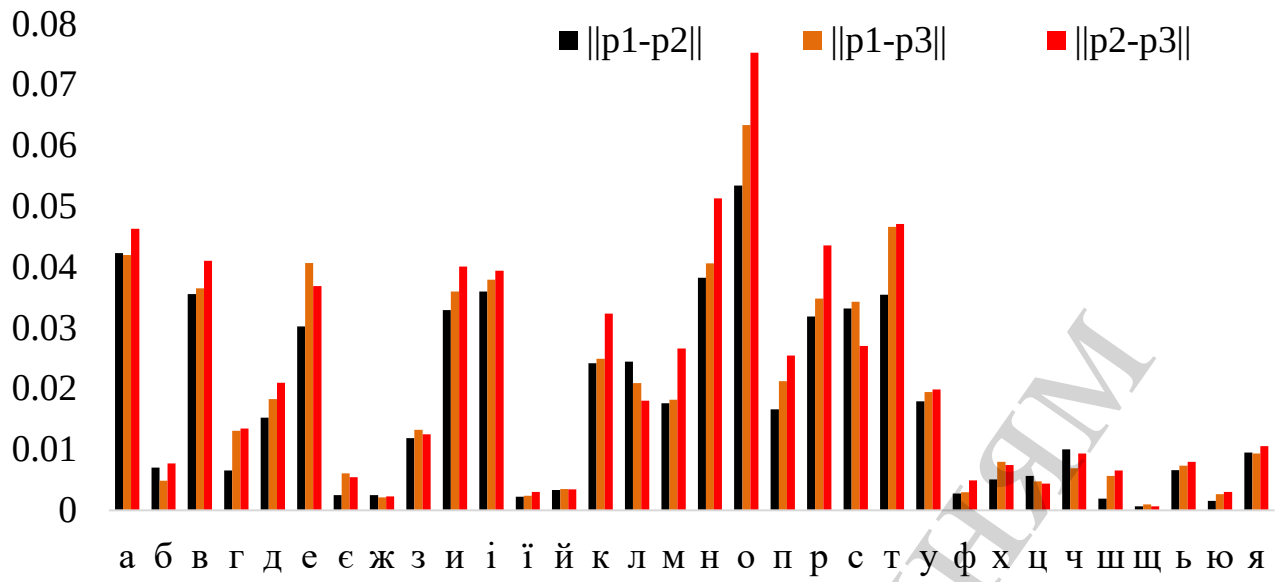


Рис. 39. Графік різниці вживання 3-грам, які починаються з конкретної літери

8. Висновки

1. Розроблено квантитативний метод ідентифікації потенційного автора тексту із множини можливих на основі порівняння результатів аналізу еталонного авторського тексту з досліджуваним. Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Його особливостями є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Наведено теоретичне та експериментальне обґрунтування 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. Розглянуто проблему реалізації визначення автора україномовного тексту із застосуванням еталонних характеристик авторського мовлення на основі методів NLP та стилеметрії. Це важливо, тому що введення інформаційних технологій стилеметрії для визначення автора текстового контенту призводить до більш високого коефіцієнта надійності визначення авторства щодо досліджуваного тексту. Але є об’єктивні труднощі, що пов’язані з точністю визначення приналежності тексту авторству конкретній людині, тому що мала вибірка серед одноосібних науково-технічних публікацій (більшість статей цієї галузі пишуть у співавторстві). Лише з врахуванням їх персональних характристик через навчання системи можна значно зменшити коло потенційних авторів конкретного наково-технічного тексту. В рамках дослідження, що описано в даній статті, розроблений квантитативний метод автоматичного визначення авторства текстового контенту на основі статистичного аналізу розподілу N-грам. Розроблена також система на основі су-

часних методів NLP та стилеметрії з врахування метрик оцінювання аналізованого тексту у порівнянні із еталонним. Також на основі сучасних методів Machine Learning розроблена система навчається уточнювати результати аналізу тексту на ступінь авторсва у порівнянні із еталоном. Це дозволяє обґрунтовано підходити до визначення якості автоматичного визначення авторства тексту науково-технічного спрямування та отримати певні ефекти від впровадження у виробництво. Зокрема, може бути уточнені коефіцієнти авторського мовлення. Одним словом, алгоритми визначення авторства на основі сучасних підходів NLP та стилеметрії у сукупності дають можливість зменшити множину потенційних авторів досліджуваного тексту. Подальший аналіз ключових слів, вживання службових слів та стійких словосполучень дозволяє більш точно визначити ступень приналежності твору конкретному автору.

3. Досліджено результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення стилю автора в україномовних наукових текстах технічного профілю. Зазвичай в системах визначення авторства використовують алгоритми визначення плагіату на метриках копірайту та рерайту. Це необхідно лише для визначення, чи робота не була запозичена повністю чи частково. Але вони не враховують ситуацію, коли робота ще не була опублікована. Квантитативний контент-аналіз текстового контенту науково-технічного спрямування використовує переваги контент-моніторингу та контент-аналізу тексту на основі методів NLP, Web-Mining та стилеметрії для визначення множини авторів, стилі мовлення яких подібні з досліджуваним уривком тексту. Це звужує коло пошуку при подальшому використанні в методах стилеметрії для визначення ступеня приналежності аналізованого тексту конкретному авторові. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв’язність мовлення, індекси винятковості та концентрації тексту. Паралельно проаналізовані такі параметри авторського стилю, як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше. Для прикладу проаналізовано 3-грами 3-х статтей. Для Статті 1 проаналізовано 78,4814 % 3-грам, для Статті 2 – 72,6332 % та для Статті 3 – 84,1271 %. Відповідно різниця вживання відповідних 3-грам між Статтями 1 та 2 є R_{12}=56,5254 %, між Статтями 2 та 3 – R_{23}=69,4271 %, між Статтями 1 та 3 – R_{13}=62.9839 %. Самі ці показники показують, що характеристики статті 1 та 2 більш подібні ($R_{23}>R_{12}$ на 12.9017 %, $R_{23} > R_{13}$ на 6.4432 %, $R_{13}> R_{12}$ на 6.4585 %, тобто $R_{23}>R_{13}>R_{12}$), ніж характеристики відповідно Статті 1–3 та 2–3. Чим менше R_{ij}, тим більша ступінь, що статті написані одним і тим же автором. Тоді в випадку Стаття 1 та 2 більш ймовірніше написана одним автором/колективом, ніж Статті 2–3 та Статті 1–3 відповідно. Стаття містить матеріали закінченого наукового дослідження в галузі інформаційних технологій в частині, що стосується комп’ютерної лінгвістки, штучного інтелекту та Machine Learning. Отримані результати, наведені у статті, дають підстави стверджувати щодо можливості втілення у реальне промислове виробництво.

Література

1. Lytvyn, V., Vysotska, V., Pukach, P., Nytrebych, Z., Demkiv, I., Kovalchuk, R., Huzyk, N. (2018). Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies, 5 (2 (95)), 16–28. doi: https://doi.org/10.15587/1729-4061.2018.142451
2. Lytvyn, V., Vysotska, V., Uhryn, D., Hrendus, M., Naum, O. (2018). Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies, 2 (2 (92)), 23–37. doi: https://doi.org/10.15587/1729-4061.2018.126009
3. Бук, С. (2008). Основи статистичної лінгвістики. Львів, 124.
4. Lytvyn, V., Vysotska, V., Pukach, P., Brodyak, O., Ugryn, D. (2017). Development of a method for determining the keywords in the slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies, 2 (2 (86)), 14–23. doi: https://doi.org/10.15587/1729-4061.2017.98750
5. Lytvyn, V., Vysotska, V., Pukach, P., Bobyk, I., Uhryn, D. (2017). Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology. Eastern-European Journal of Enterprise Technologies, 4 (2 (88)), 10–19. doi: https://doi.org/10.15587/1729-4061.2017.107512
6. Lytvyn, V., Pukach, P., Bobyk, I., Vysotska, V. (2016). The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies, 5 (2 (83)), 4–12. doi: https://doi.org/10.15587/1729-4061.2016.77174
7. Lytvyn, V., Vysotska, V., Pukach, P., Vovk, M., Ugryn, D. (2017). Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies, 3 (2 (87)), 11–17. doi: https://doi.org/10.15587/1729-4061.2017.103630
8. Khomytska, I., Teslyuk, V. (2016). The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. Advances in Intelligent Systems and Computing, 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
9. Khomytska, I., Teslyuk, V., Holovatyy, A., Morushko, O. (2018). Development of methods, models, and means for the author attribution of a text. Eastern-European Journal of Enterprise Technologies, 3 (2 (93)), 41–46. doi: https://doi.org/10.15587/1729-4061.2018.132052
10. Khomytska, I., Teslyuk, V. (2018). Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level. Advances in Intelligent Systems and Computing III, 105–118. doi: https://doi.org/10.1007/978-3-030-01069-0_8
11. Khomytska, I., Teslyuk, V. (2016). Specifics of phonostatistical structure of the scientific style in English style system. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589887

12. Khomytska, I., Teslyuk, V. (2017). Modelling of phonostatistical structures of English backlingual phoneme group in style system. 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). doi: https://doi.org/10.1109/cadsm.2017.7916144

13. Khomytska, I., Teslyuk, V. (2017). Modelling of phonostatistical structures of the colloquial and newspaper styles in english sonorant phoneme group. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098738

14. Khomytska, I., Teslyuk, V. (2018). Authorship Attribution by Differentiation of Phonostatistical Structures of Styles. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526739

15. Khomytska, I., Teslyuk, V. (2019). The Software for Authorship and Style Attribution. 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM). doi: https://doi.org/10.1109/cadsm.2019.8779346

16. Khomytska, I., Teslyuk, V. (2019). Mathematical Methods Applied for Authorship Attribution on the Phonological Level. CSIT: Proceedings of the XIVth Scientific and Technical Conference, 7–11.

17. Большакова, Е., Клышинский, Э., Ландэ, Д., Носков, А., Пескова, О., Ягунова, Е. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. Москва: МИЭМ, 272.

18. Анисимов, А., Марченко, А. (2002). Система обработки текстов на естественном языке. Искусственный интеллект, 4, 157–163.

19. Перебийніс, В. (2000). Математична лінгвістика. Українська мова. Київ, 287–302.

20. Перебийніс, В. (2013). Статистичні методи для лінгвістів. Вінниця, 176.

21. Браславский, П. И. (2006). Интеллектуальные информационные системы. Тема 7 Тематическая категоризация. URL: https://docplayer.ru/36866470-Intellektualnye-informacionnye-sistemy-tema-7-tematicheskaya-kategorizaciya-pavel-isaakovich-braslavskiy-vesenniy-semestr-2006.html

22. Ланде, Д., Жигало, В. (2008). Підхід до рішення проблем пошуку двомовного плагіату. Проблеми інформатизації та управління, 2 (24), 125–129.

23. Варфоломеев, А. (2000). Психосемантика слова и лингвостатистика текста. Калининград, 37.

24. Марченко, О. (2006). Моделювання семантичного контексту при аналізі текстів на природній мові. Вісник Київського університету, 3, 230–235.

25. Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl., 2 (6), 1930–1938.

26. Лінгвометрія. Victana. URL: http://victana.lviv.ua/nlp/linhvometriia

27. Сушко, С. О., Фомичова, Л. Я., Барсуков, Є. С. (2010). Частоти повторюваності букв і біграм у відкритих текстах українською мовою. Захист інформації, 12 (3 (48)). doi: https://doi.org/10.18372/2410-7840.12.1968

28. Когнитивная стилометрия: к постановке проблемы. URL: http://www.manekin.narod.ru/hist/styl.htm

29. Кочерган, М. (2005). Вступ до мовознавства. Київ, 368.
30. Родионова, Е. (2008). Методы атрибуции художественных текстов. Структурная и прикладная лингвистика, 7, 118–127.
31. Мещеряков, Р. В., Васюков, Н. С. (2005). Модели определения авторства текста. Измерения, автоматизация и моделирование в промышленности и научных исследованиях, 25–29.
32. Морозов, Н. А. Лингвистические спектры. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
33. Mobasher, B. (2007). Data Mining for Web Personalization. The Adaptive Web. Lecture Notes in Computer Science, 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
34. Dinucă, C. E., Ciobanu, D. (2012). Web Content Mining. Annals of the University of Petroşani. Economics, 12 (1), 85–92.
35. Xu, G., Zhang, Y., Li, L. (2010). Web Content Mining. Web Mining and Social Networking, 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4
36. Mishler, A., Crabb, E. S., Paletz, S., Hefright, B., Golonka, E. (2015). Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis. HCI International 2015 - Posters' Extended Abstracts, 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108
37. Бублейник, Л. (2000). Особливості художнього мовлення. Луцьк, 179.
38. Kowalska, K., Cai, D., Wade, S. (2012). Sentiment Analysis of Polish Texts. International Journal of Computer and Communication Engineering, 1 (1), 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12
39. Kotsyba, N. (2009). The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR). Organization and Development of Digital Lexical Resources, 55–60.
40. Lytvyn, V., Vysotska, V., Rzheuskyi, A. (2019). Technology for the Psychological Portraits Formation of Social Networks Users for the IT Specialists Recruitment Based on Big Five, NLP and Big Data Analysis. Proceedings of the 1st International Workshop on Control, Optimisation and Analytical Processing of Social Networks (COAPSN-2019), 2392, 147–171.
41. Lytvyn, V., Vysotska, V., Rusyn, B., Pohreliuk, L., Berezin, P., Naum, O. (2019). Textual Content Categorizing Technology Development Based on Ontology. Workshop Proceedings of the 8th International Conference on "Mathematics. Information Technologies. Education", 2386, 234–254.
42. Lytvyn, V., Kuchkovskiy, V., Vysotska, V., Markiv, O., Pabyrivskyy, V. (2018). Architecture of System for Content Integration and Formation Based on Cryptographic Consumer Needs. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526669
43. Berko, A., Alieksieiev, V. (2018). A Method to Solve Uncertainty Problem for Big Data Sources. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478460
44. Gozhyj, A., Kalinina, I., Vysotska, V., Gozhyj, V. (2018). The Method of Web-Resources Management Under Conditions of Uncertainty Based on Fuzzy

Logic. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526761

45. Lytvyn, V., Vysotska, V., Dosyn, D., Burov, Y. (2018). Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology, 15 (2), 66–85.

46. Nytrebych, Z. M., Malanchuk, O. M., Il'kiv, V. S., Pukach, P. Ya. (2017). Homogeneous problem with two-point conditions in time for some equations of mathematical physics. Azerbaijan Journal of Mathematics, 7 (2), 180–196.

47. Nytrebych, Z., Il'kiv, V., Pukach, P., Malanchuk, O. (2018). On nontrivial solutions of homogeneous Dirichlet problem for partial differential equations in a layer. Kragujevac Journal of Mathematics, 42 (2), 193–207. doi: https://doi.org/10.5937/kgjmath1802193n

48. Nytrebych, Z., Malanchuk, O., Il'kiv, V., Pukach, P. (2017). On the solvability of two-point in time problem for PDE. Italian Journal of Pure and Applied Mathematics, 38, 715–726.

49. Pukach, P. Ya., Kuzio, I. V., Nytrebych, Z. M., Ilkiv, V. S. (2017). Analytical methods for determining the effect of the dynamic process on the nonlinear flexural vibrations and the strength of compressed shaft. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, 5, 69–76.

50. Pukach, P. Y., Kuzio, I. V., Nytrebych, Z. M., Il'kiv, V. S. (2018). Asymptotic method for investigating resonant regimes of nonlinear bending vibrations of elastic shaft. Scientific Bulletin of National Mining University, 1, 68–73. doi: https://doi.org/10.29202/nvngu/2018-1/9

51. Nytrebych, Z., Ilkiv, V., Pukach, P., Malanchuk, O., Kohut, I., Senyk, A. (2019). Analytical method to study a mathematical model of wave processes under two-point time conditions. Eastern-European Journal of Enterprise Technologies, 1 (7 (97)), 74–83. doi: https://doi.org/10.15587/1729-4061.2019.155148

52. Pukach, P., Il'kiv, V., Nytrebych, Z., Vovk, M., Pukach, P. (2017). On the Asymptotic Methods of the Mathematical Models of Strongly Nonlinear Physical Systems. Advances in Intelligent Systems and Computing, 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30

53. Lavrenyuk, S. P., Pukach, P. Y. (2007). Mixed problem for a nonlinear hyperbolic equation in a domain unbounded with respect to space variables. Ukrainian Mathematical Journal, 59 (11), 1708–1718. doi: https://doi.org/10.1007/s11253-008-0020-0

54. Pukach, P. Y. (2016). Investigation of Bending Vibrations in Voigt–Kelvin Bars with Regard for Nonlinear Resistance Forces. Journal of Mathematical Sciences, 215 (1), 71–78. doi: https://doi.org/10.1007/s10958-016-2823-0

55. Pukach, P., Il'kiv, V., Nytrebych, Z., Vovk, M. (2017). On nonexistence of global in time solution for a mixed problem for a nonlinear evolution equation with memory generalizing the Voigt-Kelvin rheological model. Opuscula Mathematica, 37 (45), 735. doi: https://doi.org/10.7494/opmath.2017.37.5.735

56. Pukach, P. Y. (2012). On the unboundedness of a solution of the mixed problem for a nonlinear evolution equation at a finite time. Nonlinear Oscillations, 14 (3), 369–378. doi: https://doi.org/10.1007/s11072-012-0164-6

57. Pukach, P. Y. (2014). Qualitative Methods for the Investigation of a Mathematical Model of Nonlinear Vibrations of a Conveyer Belt. Journal of Mathematical Sciences, 198 (1), 31–38. doi: https://doi.org/10.1007/s10958-014-1770-x

58. Bezobrazov, S., Sachenko, A., Komar, M., Rubanau, V. (2016). The Methods of Artificial Intelligence for Malicious Applications Detection in Android OS. International Journal of Computing, 15 (3), 184–190.

59. Dunets, O., Wolff, C., Sachenko, A., Hladiy, G., Dobrotvor, I. (2017). Multi-agent system of IT project planning. 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). doi: https://doi.org/10.1109/idaacs.2017.8095141

60. Vysotska, V., Lytvyn, V., Burov, Y., Berezin, P., Emmerich, M., Basto Fernandes, V. (2019). Development of Information System for Textual Content Categorizing Based on Ontology. CEUR Workshop Proceedings, 53–70.

61. Vysotska, V., Lytvyn, V., Burov, Y., Gozhyj, A., Makara, S. (2018). The consolidated information web-resource about pharmacy networks in city. Proceedings of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018), 2255, 239–255. URL: http://ceur-ws.org/Vol-2255/paper22.pdf

62. Rusyn, B., Vysotska, V., Pohreliuk, L. (2018). Model and Architecture for Virtual Library Information System. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526679

63. Lytvyn, V., Vysotska, V., Dosyn, D., Lozynska, O., Oborska, O. (2018). Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478500

64. Lytvyn, V., Vysotska, V., Burov, Y., Bobyk, I., Ohirko, O. (2018). The Linguometric Approach for Co-authoring Author's Style Definition. 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). doi: https://doi.org/10.1109/idaacs-sws.2018.8525741

65. Zdebskyi, P., Vysotska, V., Peleshchak, R., Peleshchak, I., Demchuk, A., Krylyshyn, M. (2019). An Application Development for Recognizing of View in Order to Control the Mouse Pointer. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 55–74.

66. Veres, O., Rusyn, B., Sachenko, A., Rishnyak, I. (2018). Choosing the method of finding similar images in the reverse search system. Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Volume I: Main Conference (COLINS 2018), 2136, 99–107.

67. Vysotska, V., Lytvyn, V., Hrendus, M., Kubinska, S., Brodyak, O. (2018). Method of Textual Information Authorship Analysis Based on Stylometry. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sci-

ences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526608

68. Gozhyj, A., Chyrun, L., Kowalska-Styczen, A., Lozynska, O. (2018). Uniform Method of Operative Content Management in Web Systems. Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Systems. Volume I: Main Conference (COLINS 2018), 2136, 62–77. URL: http://ceur-ws.org/Vol-2136/10000062.pdf

69. Vysotska, V., Burov, Y., Lytvyn, V., Demchuk, A. (2018). Defining Author's Style for Plagiarism Detection in Academic Environment. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 128–133. doi: https://doi.org/10.1109/dsmp.2018.8478574

70. Chyrun, L., Vysotska, V., Kis, I., Chyrun, L. (2018). Content Analysis Method for Cut Formation of Human Psychological State. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478619

71. Gozhyj, A., Vysotska, V., Yevseyeva, I., Kalinina, I., Gozhyj, V. (2018). Web Resources Management Method Based on Intelligent Technologies. Advances in Intelligent Systems and Computing III, 206–221. doi: https://doi.org/10.1007/978-3-030-01069-0_15

72. Chyrun, L., Kis, I., Vysotska, V., Chyrun, L. (2018). Content Monitoring Method for Cut Formation of Person Psychological State in Social Scoring. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2018.8526624

73. Demchuk, A., Lytvyn, V., Vysotska, V., Dilai, M. (2019). Methods and Means of Web Content Personalization for Commercial Information Products Distribution. Lecture Notes in Computational Intelligence and Decision Making, 332–347. doi: https://doi.org/10.1007/978-3-030-26474-1_24

74. Lytvyn, V., Vysotska, V., Kuchkovskiy, V., Bobyk, I., Malanchuk, O., Ryshkovets, Y. et. al. (2019). Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies, 1 (2 (97)), 18–39. doi: https://doi.org/10.15587/1729-4061.2019.154709

75. Vysotska, V., Fernandes, V. B., Lytvyn, V., Emmerich, M., Hrendus, M. (2018). Method for Determining Linguometric Coefficient Dynamics of Ukrainian Text Content Authorship. Advances in Intelligent Systems and Computing III, 132–151. doi: https://doi.org/10.1007/978-3-030-01069-0_10

76. Kravets, P. (2010). The control agent with fuzzy logic. Perspective Technologies and Methods in MEMS Design, 40–41.

77. Kravets, P. (2007). The Game Method for Orthonormal Systems Construction. 2007 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics. doi: https://doi.org/10.1109/cadsm.2007.4297555

78. Kravets, P. (2016). Game Model of Dragonfly Animat Self-Learning. Perspective Technologies and Methods in MEMS Design, 195–201.

79. Basyuk, T. (2015). The main reasons of attendance falling of internet resource. 2015 Xth International Scientific and Technical Conference “Computer Sciences and Information Technologies” (CSIT). doi: https://doi.org/10.1109/stc-csit.2015.7325440

80. Chyrun, L., Kowalska-Styczen, A., Burov, Y., Berko, A., Vasevych, A., Pelekh, I., Ryshkovets, Y. (2019). Heterogeneous Data with Agreed Content Aggregation System Development. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 2386, 35–54.

81. Chyrun, L., Burov, Y., Rusyn, B., Pohreliuk, L., Oleshek, O. et. al. (2019). Web Resource Changes Monitoring System Development. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 2386, 255–273.

82. Vysotska, V., Burov, Y., Lytvyn, V., Oleshek, O. (2019). Automated Monitoring of Changes in Web Resources. Lecture Notes in Computational Intelligence and Decision Making, 348–363. doi: https://doi.org/10.1007/978-3-030-26474-1_25

83. Chyrun, L., Gozhyj, A., Yevseyeva, I., Dosyn, D., Tyhonov, V., Zakharchuk, M. (2019). Web Content Monitoring System Development. Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference, 2362, 126–142.

84. Rzheuskyi, A., Gozhyj, A., Stefanchuk, A., Oborska, O., Chyrun, L., Lozynska, O. et. al. (2019). Development of Mobile Application for Choreographic Productions Creation and Visualization. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 2386, 340–358.

85. Lytvynenko, V., Savina, N., Krejci, J., Voronenko, M., Yakobchuk, M., Kryvoruchko, O. (2019). Bayesian Networks' Development Based on Noisy-MAX Nodes for Modeling Investment Processes in Transport. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 2386, 1–10.

86. Lytvynenko, V., Lurie, I., Krejci, J., Voronenko, M., Savina, N., Taif, M. A. (2019). Two Step Density-Based Object-Inductive Clustering Algorithm. Workshop Proceedings of the 8th International Conference on “Mathematics. Information Technologies. Education”, 2386, 117–135.

87. Antonyuk, N., Medykovskyy, M., Chyrun, L., Dverii, M., Oborska, O., Krylyshyn, M. et. al. (2019). Online Tourism System Development for Searching and Planning Trips with User’s Requirements. Advances in Intelligent Systems and Computing, 831–863. doi: https://doi.org/10.1007/978-3-030-33695-0_55

88. Rzheuskyi, A., Kutyuk, O., Voloshyn, O., Kowalska-Styczen, A., Voloshyn, V., Chyrun, L. et. al. (2019). The Intellectual System Development of Distant Competencies Analyzing for IT Recruitment. Advances in Intelligent Systems and Computing, 696–720. doi: https://doi.org/10.1007/978-3-030-33695-0_47

89. Rusyn, B., Pohreliuk, L., Rzheuskyi, A., Kubik, R., Ryshkovets, Y., Chyrun, L. et. al. (2019). The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts’ Bonfires. Advances in Intelligent Systems and Computing, 734–756. doi: https://doi.org/10.1007/978-3-030-33695-0_49

90. Chyrun, L., Leshchynskyy, E., Lytvyn, V., Rzheuskyi, A.,Vysotska, V., Borzov, Y. (2019). Intellectual Analysis of Making Decisions Tree in Information Systems of Screening Observation for Immunological Patients. CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), 281–296. URL: http://ceur-ws.org/Vol-2488/paper25.pdf

91. Lytvyn, V., Burov, Y., Kravets, P., Vysotska, V., Demchuk, A., Berko, A. et. al. (2019). Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification. CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), 354–368. URL: http://ceur-ws.org/Vol-2488/paper31.pdf

92. Antonyuk, N., Chyrun, L., Andrunyk, V., Vasevych, A., Chyrun, S., Gozhyj, A. et. al. (2019). Medical News Aggregation and Ranking of Taking into Account the User Needs. CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), 369–382. URL: http://ceur-ws.org/Vol-2488/paper32.pdf

93. Babichev, S., Taif, M. A., Lytvynenko, V., Osypenko, V. (2017). Criterial analysis of gene expression sequences to create the objective clustering inductive technology. 2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO). doi: https://doi.org/10.1109/elnano.2017.7939756

94. Babichev, S., Korobchynskyi, M., Lahodynskyi, O., Korchomnyi, O., Basanets, V., Borynskyi, V. (2018). Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles. Eastern-European Journal of Enterprise Technologies, 1 (4 (91)), 19–32. doi: https://doi.org/10.15587/1729-4061.2018.123634

95. Babichev, S., Lytvynenko, V., Osypenko, V. (2017). Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098832

96. Babichev, S. A., Gozhyj, A., Kornelyuk, A. I., Lytvynenko, V. I. (2017). Objective clustering inductive technology of gene expression profiles based on sota clustering algorithm. Biopolymers and Cell, 33 (5), 379–392. doi: https://doi.org/10.7124/bc.000961

97. Pasichnyk, V., Shestakevych, T. (2016). The Model of Data Analysis of the Psychophysiological Survey Results. Advances in Intelligent Systems and Computing, 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18

98. Zhezhnych, P., Markiv, O. (2017). Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects. Advances in Intelligent Systems and Computing, 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45

99. Davydov, M., Lozynska, O. (2017). Information system for translation into ukrainian sign language on mobile devices. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098734

100. Davydov, M., Lozynska, O. (2017). Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies. Advances in Intelligent Systems and Computing, 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7

101. Davydov, M., Lozynska, O. (2016). Linguistic models of assistive computer technologies for cognition and communication. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589898

102. Vysotska, V., Chyrun, L. (2015). Methods of information resources processing in electronic content commerce systems. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.

103. Andrunyk, V., Chyrun, L., Vysotska, V. (2015). Electronic content commerce system development. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.

104. Alieksieieva, K., Berko, A., Vysotska, V. (2015). Technology of commercial web-resource processing. Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February.

105. Mykich, K., Burov, Y. (2016). Uncertainty in situational awareness systems. 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). doi: https://doi.org/10.1109/tcset.2016.7452165

106. Mykich, K., Burov, Y. (2016). Algebraic Framework for Knowledge Processing in Systems with Situational Awareness. Advances in Intelligent Systems and Computing, 217–227. doi: https://doi.org/10.1007/978-3-319-45991-2_14

107. Mykich, K., Burov, Y. (2016). Research of uncertainties in situational awareness systems and methods of their processing. Eastern-European Journal of Enterprise Technologies, 1 (4 (79)), 19–27. doi: https://doi.org/10.15587/1729-4061.2016.60828

108. Mykich, K., Burov, Y. (2016). Algebraic model for knowledge representation in situational awareness systems. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2016.7589896