УДК 004.89
DOI: 10.15587/1729-4061.2018.142451

Розроблення лінгвометричного методу автоматичного визначення автора текстового контенту на основі статистичного аналізу коефіцієнтів мовної різноманітності

В. В. Литвин, В. А. Висоцька, П. Я. Пукач, З. М. Нитребич, І. І. Демків, Р. А. Ковальчук, Н. М. Гузик

Розробленно лінгвометричний метод алгоритмічного забезпечення процесів контент-моніторінгу для розв'язання задачі автоматичного визначення автора україномовного текстового контенту на основі технології статистичного аналізу коефіцієнтів мовної різноманітності. 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. Проаналізовані також 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.

Особливостями розробленого є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього провадився аналіз флексій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації.

Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури

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

1. Вступ

Важливими завданнями мовознавства на основі лінґвометрії є створення і порівняння словників (у тому числі частотних та статистичних), автоматичних словників, тезаурусів, систем стенографії, автоматичне визначення мови, інфор-

маційний пошук тощо [1]. Наприклад, для моделювання процесів інформаційного пошуку знаходять статистичні і перехідні ймовірності морфем тексту [2]. На основі побудованих таблиць моделюють перевірку досліджуваного слова на наявність помилки, пропонують кілька найбільш ймовірних варіантів [3]. У свою чергу стилеметрія як підрозділ прикладної лінґвістики виявляє та аналізує кількісні характеристики певного функціонального стилю мови чи мовлення авторів текстового контенту, тобто авторської атрибуції [4]. Атрибуція полягає у визначенні методома квантитативної лінгвістики достовірності, автентичності авторського твору, його автора, місця й часу створення на основі аналізу технологічних і стилістичних закономірностей та особливостей коефіцієнтів мовної різноманітності конкретного авторач і/або конкрентного текстового твору [5]. Наприклад, однією із відомих мовознавчих проблем є процес визначення авторської атрибуції уривків певного текстового контенту [6]. Для цього обчислють частоти слововживань у запропонованих уривках [7]. Використовуючи частотні словники авторської творчості загалом чи окремих його творів, визначають автора твору (або твір – якщо це дозволяє словник) [8]. Недоліком є збереження або автоматичне генерування великих масивів даних у вигляді частотних словників авторських творів [9]. Опрацювання таких словників вимагає багато часу, а збереження – багато ресурсів [10]. У свою чергу, є автори з малочисельною творчістю, що унеможливлю точне відворення результатів аналізу авторської атрибуції [11]. Відомий метод датування для визначення тривалості роздільного існування двох споріднених мов, ґрунтується на припущенні про те, що основна частина лексичного складу будь–якої мови (ядерна лексика) змінюється з однаковою швидкістю і вимагає підрахунку процентного співвідношення спільних елементів у основному словнику [12]. Модифіковані методи глотохронології застосовують для визначення динаміки зміни авторскього мовлення в його текстовому контенті на протязі тривалого часу для датування наближеного періоду, в якому був створений конкртниий текст твору цього автора [13]. Тому задача автоматичного визначення автора текстового контенту є актуальною й потребує нових (досконаліших) підходів до її розв'язування, наприклад, на основі статистичного аналізу коефіцієнтів мовної різноманітності [14].

2. Аналіз літературних даних і постановка проблеми

Вагомим значенням у квантитативній лінгвостатистиці є розподіл (дистрибуція) лінгвістичної одиниці у тексті – присутність лінгвістичної одиниці в різних (зазвичай рівних) підвибірках (уривках) [15]. Якщо досліджувана лінгвістична одиниця функціонує тільки в одній нідвибірці, хоча й з високою частотою, то така вибірка є нерепрезентативною стосовно цієї лінгвістичної одиниці [16]. Важливо, коли досліджувана лінгвістична одиниця є рівномірно розподіленою в генеральній сукупності [17]. Для цього аналізують коефіцієнт розповсюдженості [18]: $K_r=N_p/N_z$, де N_p – відношення кількості підвибірок з певною лінгвістичною одиницею, N_z – загальна кількість підвибірок. Проте характеристики, одержані на матеріалі вибірки, зазвичай відрізнятються від реальних характеристик генеральної сукупності, оскільки завжди присутня в квантитативній лінгвостатистиці відносна неточність дослідження [19]. Розподіл частоти

лінгвістичних одиниць мови в текстовому контенті має певну регулярність і утворює його статистичну (частотну, ймовірнісну) структуру [20]. Такий розподіл є відмінним для кожної з мовних елементів – лексем, морфем, фонем тощо [21]. Тому лінгвостатистичні параметри авторських стилів, встановлені на різних рівнях (фонемних, морфемних, *N*-грамних, лексемних тощо), мають неоднакову стилеідентифікованму потужність автоського мовлення для різних пар стилів [22]. Наприклад, споріднені стилі чіткіше розмежовані на синтаксичному рівні, а менш споріднені – на лексичному [23]. Для цього автоматично створюють частотні словники певних лінгвістичних одиниць та завдяки ним аналізують середню повторюваність слова в тексті, коефіцієнт hapax legomena (слова, які мають частоту 1 у досліджуваній вибірці), індекс винятковості, індекс концентрації тощо [1–5, 14, 24].

За даними ЧС обчислюють такі характеристики як багатство словника, *індекс різноманітності* (K_l) – відношення обсягу словника лексем (W) до обсягу тексту (N), тобто $K_l = W / N$. Згідно табл. 1 найрізноманітніша, найбагатша лексика – у поезії, далі за спадом – у художній прозі, розмовно–побутовому стилі, публіцистиці, науковому та офіційно–діловому стилі [14, 25].

Таблиця 1

Результати коефіцієнтів мовлення згідно стилів української мови [14]

Стиль	W/N	W_1/N	W_1/W	W_{10}/W	W_{10t}/N
науковий	0,059	0,427	0,025	0,189	0,890
публіцистичний	0,070	0,450	0,031	0,121	0,804
діловий	0,030	0,280	0,0085	0,303	0,935
поетичний	0,103	0,495	0,052	0,098	0,789
художньої прози	0,067	0,430	0,029	0,149	0,821
розмовний	0,073	0,465	0,034	0,161	0,789

Середня повторюваність слова у тексті A є відношенням обсягу тексту N до обсягу словника лексем W (обернена до індексу різноманітності), тобто $A = N / W$ [26]. За даними ЧС, кожне слово у розмовно–побутовому стилі в середньому вжито 14 разів, а в науковому стилі – 17 [27].

Індекс винятковості храктеризує варіативність лексики, тобто частку тексту (словника), яку займають слова, що трапилися 1 раз (табл. 1) [28]:

– *словника* I_{wt} – відношення кількості лексем із частотою 1 W_1 до загальної кількості лексем: $I_{wt}=W_1/W$ [14];

– *тексту* I_t відношення кількості лексем із частотою 1 W_1 до обсягу тексту N: $I_t=W_1/N$ [14].

Індекс концентрації вказує на частку тексту (словника), яку займають слова, що трапилися 10 разів і більше (табл. 1) [29]:

– *словника* I_{kt} – відношення кількості слів у словнику з абсолютною частотою 10 і більше (W_{10}) до загальної кількості слів у словнику (W): $I_{kt} = W_{10}/W$ [14];

– *тексту* I_{tn} – відношення суми абсолютних частот слів з абсолютною частотою 10 і більше W_{10t} до обсягу тексту N: $I_{tn}=W_{10t}/N$ [14].

Як видно із ЧС, мовлення надає перевагу невеликій кількості одиниць, які часто використовують [30]. Формують ядро будь–якої мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними [31]. Цю закономірність зауважив ще учений Дьюї на поч. XX ст., назвавши її *законом переваги* [32]. Детальніше дослідив цю закономірність німецький мовознавець Дж. Ціпф, сформулювавши закон *Zipf's law*, який встановлює залежності [33]:

– частоти слова та його рангу у словнику: чим частотніше слово, тим вищий його ранг при $F \times i = const$, де F – частота слова в частотному словнику, i – ранг цього слова [34];

– частоти слова та його довжини: чим частотніше слово, тим воно коротше при $k=C \lg r$, де k – довжина слова у фонемах, C – стала, r – ранг [35];

– частоти слова та кількості його значень: чим частотніше слово, тим воно багатозначніше при $m = C\sqrt{f}$, де m – кількість значень слова, C – стала, f – частота слова [36];

– частоти слова та його походження: чим давніше слово, тим воно частотніше [37].

Згідно закону німецького мовознавця П. Менцерата довжина мовної конструкції (слова, словосполучення, надфразової єдності, речення) обернено пропорційна до довжини її складових (складів, слів, словосполучень і т. д.), тобто чим довша мовна конструкція, тим коротші її складові [14]. Згідно досліджень Ґ. Альтманна $y=ax^b$, де y – середня довжина складових, x – довжина мовної конструкції, b – показник, що характеризує динаміку зміни довжини складників (закон діє, якщо $b<0$) [38].

Закон Крилова встановлює залежність між кількістю багатозначних слів та частотою:

$$p_x = 1/2^x, \;\; px = (\omega - 1)^{x-1} / \omega^x,$$

де p_x – ймовірність використання слова, яке має x значень, ω – середня кількість значень слова у словнику [14].

Деякі основні кількісні характеристики мови дуже прості. Наприклад, різниця між кількістю слів (10^4–10^5), кількістю морфем (декілька тисяч), кількістю складів (від декількох сотень до декількох тисяч) і кількістю фонем (від 10 до 80) [3149]. Висловлюють припущення, що такі співвідношення пов'язані із властивістю людської пам'яті [39]. Зазначимо також, що чим частотніше слово, тим швидше людина його зможе пригадати [40]. Однак відсутні дослідження в галузі залежності змін коефіцієнтів лексичного авторського мовлення на протязі періоду його творчості [41].

3. Мета і завдання дослідження

Метою роботи є розроблення методу визначення автора у україномовних текстах на основі технології лінгвометрії.

Для досягнення мети були поставлені такі завдання:

– на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному тексті розробити алгоритми визначення автора тексту;

– розробити програмне забезпечення контент-моніторінгу для визначення автора в україномовних текстах на основі лінгвометричного аналізу визначених стопових слів текстового контенту;

– здійснити аналіз результатів експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю.

4. Метод визначення стилю автора текстового контенту

Лінґвометрія – галузь прикладної лінґвістики, що виявляє, вимірює та аналізує кількісні характеристики одиниць різних рівнів мови чи мовлення [42]. Одним зі способів охарактеризувати літературне багатство тексту є оцінювання характеру використання мовних одиниць на всіх мовних рівнях [43]. Це дає змогу ототожнювати поняття *багатство* і *різноманітність* мовлення [44]. Розрахунок коефіцієнтів мовної різноманітності повинен припускати взаємозв'язок таких коефіцієнтів, як *лексична різноманітність*, *ступінь* (*міра*) *синтаксичної складності* [14], *зв'язність мовлення*, *індекси винятковості та концентрації тексту* [45]. Оскільки коефіцієнт – величина абсолютна, можна у певних межах нехтувати довжиною порівнюваних текстів [46]. Теоретичний інтерес складає дослідження внутрішньої «динаміки» тексту в частині співставлення коефіцієнтів з різних його ділянок між собою та із загальним для всього тексту коефіцієнтом (табл. 2) [47]:

– для лексичної різноманітності чим більшим є отримуваний десятковий дріб, тим вищою є лексична різноманітність досліджуваного тексту [48];

– для синтаксичної складності чим більшим є дріб (в межах [0; 1]), тим багатослівнішими загалом є речення такого тексту, а отже, – вища можливість різноманітності синтаксичних відношень між словами в окремому реченні [49];

– для зв'язності мовлення дорівнює одиниці, коли в одному реченні є три сполучні елементи (прийменники і сполучники) [50].

Таблиця 2

Коефіцієнти різноманітності тексту [1–5, 14, 41, 47]

Коефіцієнт	Визначення	Формула	Пояснення
Лексична різноманітність	Відношення кількості слів до загальної кількості словоформ тексту. Значення коефіцієнта лежить у межах [0;1]	$K_l=W/N$	K_l – коефіцієнт лексичної різноманітності, W – кількість слів у певному тексті, N – загальна кількість слів цього тексту
Синтаксична складність	Відношення кількості речень до кількості слів певного тексту	$K_s=1–P/W$	K_s – коефіцієнт синтаксичної складності, P – кількість речень, W – кількість слів у всьому

Коефіцієнт	Визначення	Формула	Пояснення
			тексті
Коефіцієнт зв'язності мовлення	Відношення кількості прийменників і сполучників до кількості окремих речень	$K_z=(Z+S)/(3P)$	Z – кількість прийменників, S – кількість сполучників, P – кількість окремих речень
Індекс винятковості	Варіативність лексики, тобто частка тексту, яку займають слова, що трапилися 1 раз	$I_{wt}=W_1/W$	I_{wt} – індекс винятковості тексту, W_1 – кількість слів із частотою 1, W – кількість слів у всьому тексті
Індекс концентрації	Частка тексту, яку займають слова, що трапилися 10 разів і більше	$I_{kt}=W_{10}/W$	I_{kt} – індекс концентрації тексту, W_{10} – кількість слів із частотою 10 та більше, W – кількість слів у всьому тексті

Виявлено [14], що текст україномовної казки має K_z=0,77, а текст україномовної наукової статті – 3,0, тобто зв'язність у другому тексті у 3,9 разів сильніша, ніж у першому. Офіційних стандартів для коефіцієнтів різноманітності мовлення для K_l та K_s не існує [51], але орієнтиром для співставлення та оцінювання якогось тексту в однорідній групі текстів є середньостатистична норма величини коефіцієнта для рівних за довжиною уривків [52]. Мінімальним розміром (довжиною) уривка приймемо 100 слів, вважатимемо, що коефіцієнти тут уже стабілізуються, відображаючи реальні особливості мови автора [53]. Близькість або віддаленість окремого індивідуального коефіцієнта від середнього служить основою для оцінювання різноманітності мовлення у відповідному тексті [54]. Задовільними вважаються тексти, коефіцієнти різноманітності яких потрапляють у зону середніх квадратичних відхилень D від певного середнього

$$D = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - (\overline{x})^2}$$ [14, 55].

Аналіз та інтерпретація на лінґвістичному рівні стилістичних особливостей і закономірностей письменницького стилю певного автора (або певної літературної епохи) включає подані найосновніші етапи, подані в алг. 1 [14, 56].

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

*Етап 1. Відбір та первинне опрацювання текстового контенту***.** Для відбору будують фільтри тексту за параметрами (основна мова тексту, обсяг текс-

тової вибірки, часовий проміжок публікації, джерео публікації, формат тощо) [41, 57]. Основними кроками первинного опрацювання тексту є:

– приведення його до єдиного формату (наприклад усунення тегів, якщо попередня публікація є у Інтернет–ресурсі у виглдяі статичної сторінки);

– усунення інформаційного шуму (рисунків, формул, список літеру тари, анотації іншими мовами тощо), який не впливає на результат, але збільшує час опрацювання;

– приведення до єдиного обсягу (скорочення у разі потреби, забираючи неінформативні ділянки початку та закінчення тексту).

Етап 2. Лематизація текстових лінгвістичних одиниць. Об'єднання словоформ під лемою мови [14, 58].

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

Етап 4. Побудова системи частотних словників, організація на основі статистичних розподілів у потрібних частотних шкалах. Частотний словник – тип словника, де наведено кількість вживань (частоту) певної лінгвістичної одиниці мови (складу, слова, словоформи, словосполучення, ідіоми, фразеологізму) в різних текстах певного обсягу [59]. Зазвичай, подають абсолютну та відносну частоту вживання мовних одиниць, словникові статті розміщують за спаданням частот [60].

Етап 5. Пошук параметрів, що адекватно відображають структуру частотного словника. Такі параметри дають змогу сформулювати кілька основних лінгвостатистичних методів дослідження тексту [61]:

– метод опорних слів (підрахунок загальної частоти вживання та знаходження відсоткового складу службових слів [62]: прийменників, сполучників, часток);

– метод розділових знаків (підрахунок лише кількості внутрішніх і зовнішніх розділових знаків) [63];

– метод слів (підрахунок лише слів певної довжини) [64];

– метод речень (підрахунок лише речень визначеної довжини) [65];

– синтаксичний метод (підрахунок розділових знаків, слів і речень певної довжини) [66];

– комбінований (поєднання синтаксичного методу і опорних слів) [67].

Етап 6. Перевірка параметрів на ефективність. Аналіз та порівння отримани результатів на відомих авторських творах для визначення закономірностей впливу авторської стилістики на формування авторської структури частотного словника за цими параметрами [68].

Етап 7. Математичне моделювання лексикостатистичних розподілів [69].

Етап 8. Побудова статистичних класифікацій, тобто авторських еталонів, що відображають стилістичні закономірності в межах творів певного автора чи певної літературного стилю з врахуванням літературної епохи та особливостей мови, на якій написані самі аналізовані твори [70].

Етап 9. Інтерпретація результатів із позицій стилістичних уявлень у визначеному часовому проміжку, загальної й авторської стилістики з врахуванням часових параметрів [71]. Таким чином також вирішемо завдання авторської атрибуції, яке сформулюємо наступним чином. Нехай існує статистично опрацьований доробок автора (еталон). Необхідно оцінити належність певних уривків до еталону із застосуванням відповідних методів. Графічне зображення відносної частоти появи службових слів в Уривку 4 та в еталоні подане на рис. 1. Коефіцієнт кореляції для службових слів у цьому випадку складає R_{e-U4}=0,7326. Наведемо також коефіцієнти кореляції для кожного зі службових слів для уривків 1–4 (табл. 4). Аналізуючи коефіцієнти кореляції для службових слів, приходимо до висновку, що ймовірність належності уривків до досліджуваного еталону найбільшою є для Уривку 4, за ним – Уривок 2, Уривок 1, Уривок 3. Зауважимо, що для всіх чотирьох уривків простежуються стабільно високі коефіцієнти кореляції для часток, що можемо розуміти як відсутність впливу часток на авторський стиль. Додатково для уривків проаналізуємо частотності поява лише прийменників і сполучників, знайдемо відповідні коефіцієнти кореляції та порівняємо результати (табл. 3).

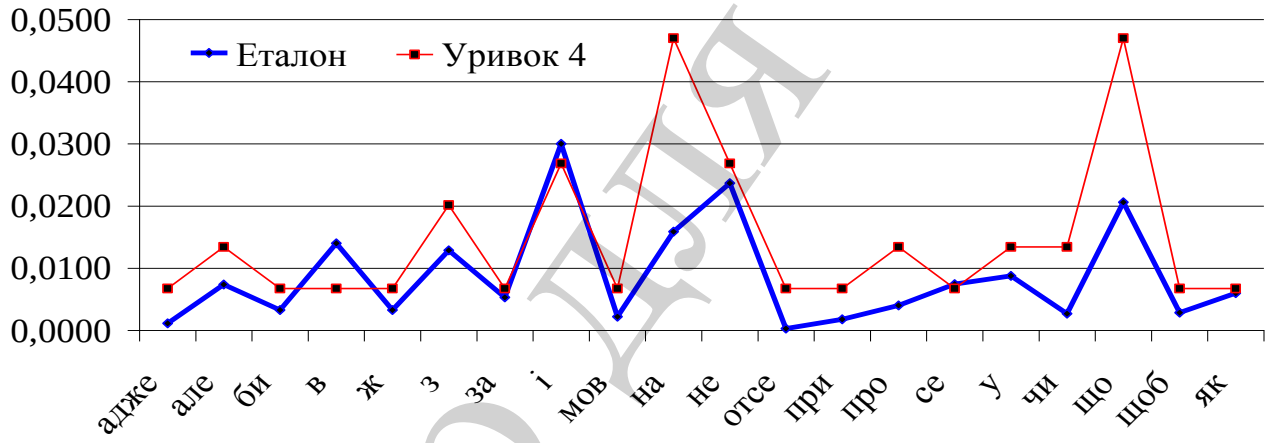


Рис. 1. Відносна частота появи службових слів в Уривку 4 та в еталоні

Таблиця 3

Коефіцієнти кореляції для службової частини мови та кожного з уривків

Уривок	Прийменник	Сполучник	Частка	R_{e-U}	R_{e-U}
1	0,72	0,79	1	0,6076	0,6900
2	0,4928	0,5714	0,9580	0,7066	0,4913
3	0,1517	0,1624	0,8800	0,2810	0,2254
4	0,5639	0,9544	0,9594	0,7326	0,6905

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

5. Результати досліджень визначення автора в україномовних науково-технічних текстах

Під час дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі Victana [25] (рис. 2). Аналізуючи складові формул для оцінки багатства твору, приходимо до висновку, що треба знайти такі величини як кількість слів і словоформ, речень, сполучників і прийменників, слів із частотою 1 та не меншою за 10. На сервері після запуску процесу розрахунку коефіцієнтів різноманітності тексту запускається алгоритм аналізу цього тексту (алг. 2).

Алгоритм 2. Аналізу стилю автоського мовлення

Етап 1. Перевірка довжини тексту – лишнє відсікається.

Етап 2. Очищення досліджуваного тексту (цифри, спецсимволи, формули, рисунки).

Етап 3. Визначення кількості речень P.

Етап 4. Визначення загальної кількості слів у тексті N.

Етап 5. Визначення кількості слів W (за частотним словником основ слів).

Етап 6. Розрахунок коефіцієнта лексичної різноманітності: $K_l=W/N$.

Етап 7. Розрахунок коефіцієнта синтаксичної складності: $K_s=1–P/W$.

Етап 8. Визначення кількості слів, що зустрілися точно один раз, тобто W_1.

Етап 9. Розрахунок індексу винятковості тексту : $I_{wt}=W_1/W$.

Етап 10. Визначення кількості слів, що зустрілися більше 9 разів, тобто W_{10}.

Етап 11. Розрахунок індексу концентрації тексту: $I_{kt}=W_{10}/W$.

Етап 12. Визначення кількості прийменників Z.

Етап 13. Визначення кількості сполучників S.

Етап 14. Розрахунок коефіцієнта зв’язності мовлення: $K_z=(Z+S)/(3*P)$.

Етап 15. Виведення результатів на Web сторінку сайту Victana [25].

Аналізуючи складові формул для оцінення багатства твору, бачимо, що треба знайти кількість речень, слів і словоформ, прийменників і сполучників, слів із частотою 1 та частотою, не меншою за 10. Для зручності внесемо знайдені дані у таблицю. На інформаційному ресурсі передається сформована таблиця (табл. 4) та отримані результати дослідження виводяться на екран.

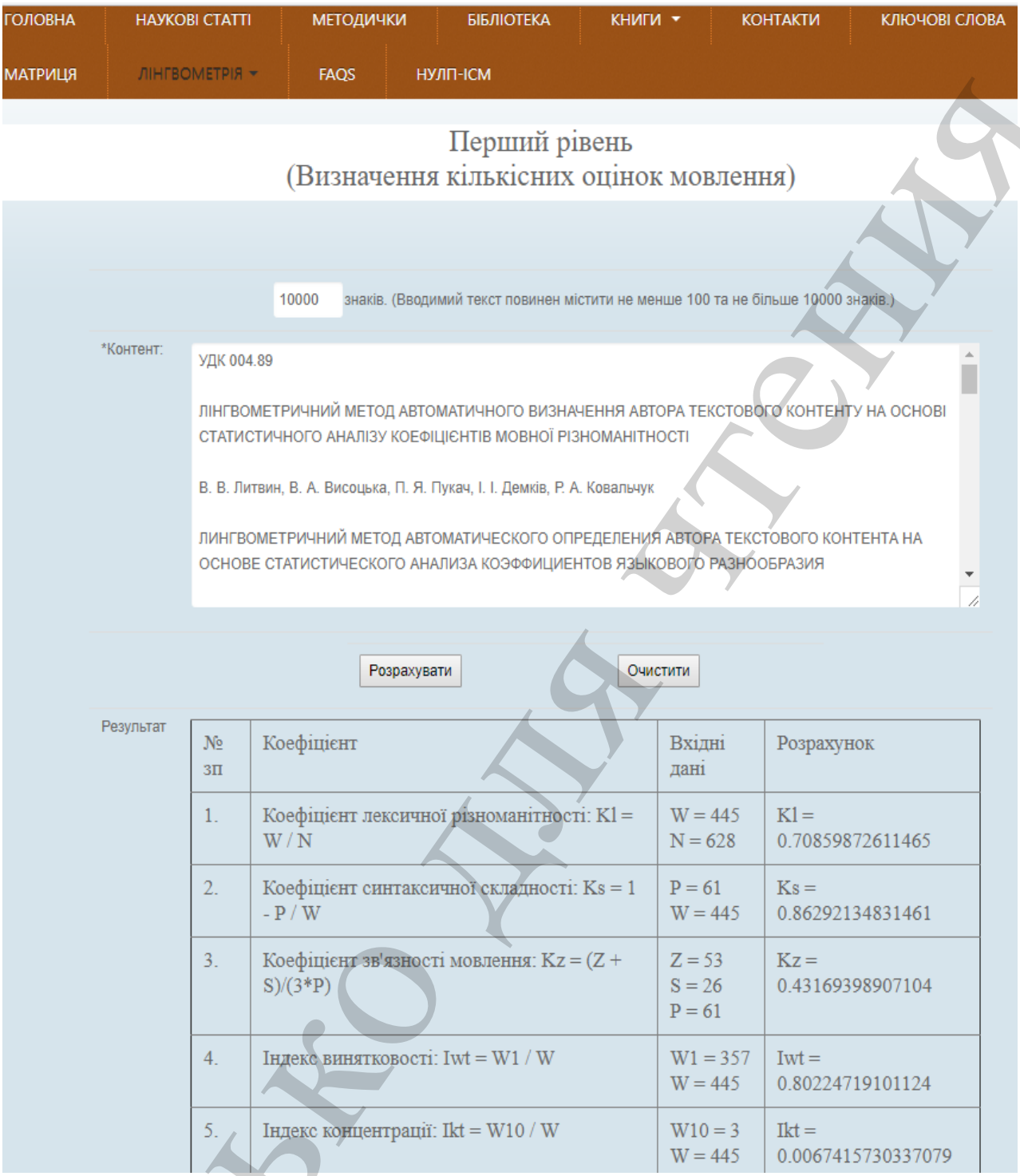


Рис. 2. Результат роботи алгоритму на Web-ресурсі Victana [25]

Таблиця 4

Приклад сформованої таблиці як результат роботи алгоритму аналізу стилю автора публікації на інформаційному ресурсі Victana [25]

Коефіцієнт	Дані	Розрахунок
лексичної різноманітності: $K_l=W/N$	W=184; N=295	K_l=0.6237
синтаксичної складності: $K_s=1–P/W$	P=18; W=184	K_s=0.902
зв’язності мовлення: $K_z=(Z+S)/(3*P)$	Z=20; S=28; P=18	K_z=0.889

винятковості: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0.7663
концентрації: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0.01

Спираючись на викладене вище, оцінимо багатство уривків творів одноосібних наукових статей технічного спрямування Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2017 рр. [25] за допомогою коефіцієнтів різноманітності та зв'язності мовлення, індексів винятковості та концентрації тексту. Для аналізу виберемо частину першу (10000 знаків) кожної статті (алг. 3).

Алгоритм 3. Аналіз статистики функціонування системи виявлення множини стопових слів із 215 наукових статей технічного спрямування

Етап 1. Аналіз 100 наукових статей на визначення діапазону оптимального розміру досліджуваного тексту. Спочатку були проаналізовані тексти в повному обсязі, а потім ці тексти були проаналізовані на різні величини знаків. Результати показали, що оптимальним дослідженням текстів є діапазон [100;10000] знаків. Менше 100 знаків – неінформативна отримана інформація, часто значення коефіцієнтів різних авторів подібні, а одного ж автора на різних тестах – суттєво різняться. Якщо більше 10000 знаків – суттєво коефіцієнти вже не змінюються, але аналоги для дослідження мають різну довжину і з–за браку різноманітності аналогів великої довжини, було обрано максимальне число для аналізу 10000.

Етап 2. Аналіз понад 200 одноосібних робіт технічного спрямування понад 50 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 3. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 4. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення стилів мовлення цих авторів.

Етап 5. Аналіз отриманих коефіцієнтів мовлення біля 100 різних авторів за період 2001–2017 рр. для визначення підмножини авторів з подібним стилем, що і 4 еталоні роботи (колективні роботи, автори яких присутні серед досліджуваних одноосібних робіт).

Етап 6. Аналіз отриманих результатів на етапі 5. Перевірити, чи в отриманих підмножинах присутні справжні автори цих еталоних текстів. Обрати найкращий алгоритм для визначення стилю автора в україномовних науково–технічних текстах на основі технології квантитативної лінвостатистики

Для чистоти дослідження необхідно проаналізувати, чи впливає час публікації робіт на коефіцієнти різноманітності тексту, тобто чи не змінюються ці коефіцієнти з часом на вибірці тих самих авторів та текстів. Спочатку проаналізуємо як змінюється загальний обсяг слів в однакових за розміром уривках в діапазоні 2001–2017 рр. Як бачимо з часом ті ж самі автори частіше вживають коротші слова (рис. 3, *а*).

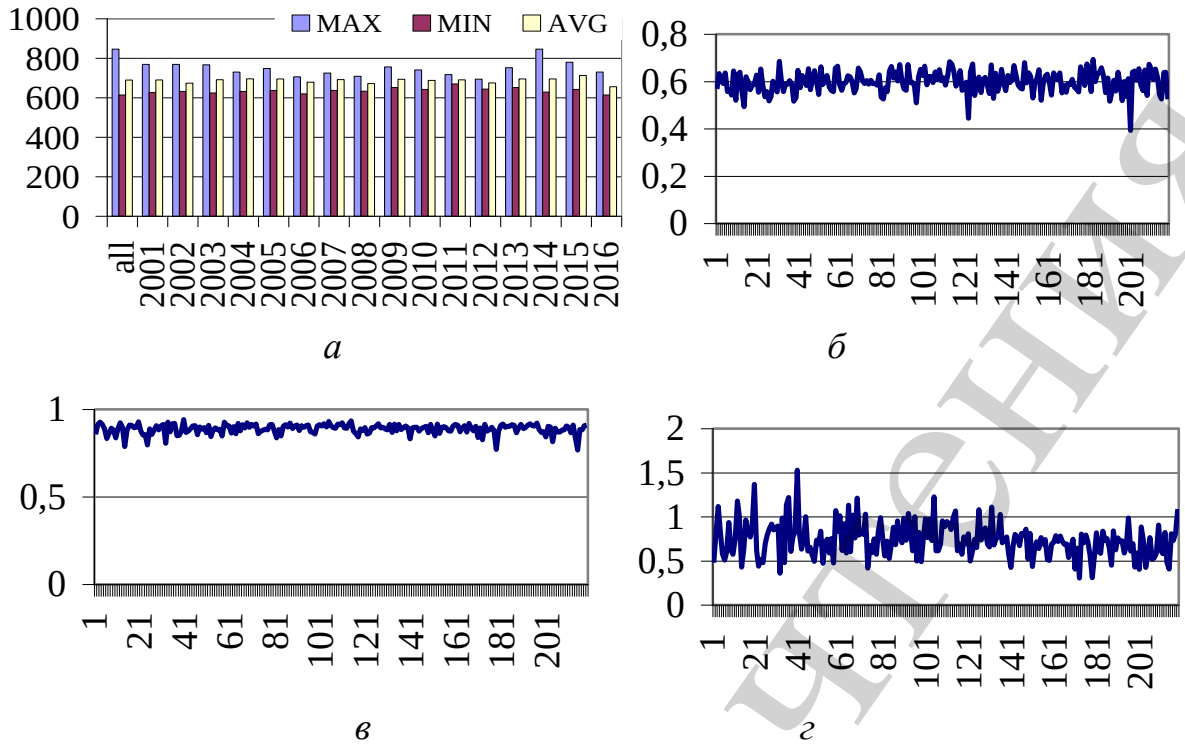


Рис. 3. Розподіл: *а* – слів та коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: *б* – K_l; *в* – K_s; *г* – K_z

З часом коефіцієнт лексичної різноманітності K_l суттєво не змінюється (рис. 3, *б–г*). Аналогічно з часом коефіцієнт синтаксичної складності K_s також суттєво не змінюється. А ось коефіцієнт зв'язності мовлення K_z з часом за 16 років зменшується, хоча не суттєво. На початках (2001 р.) коливається в діапазоні [0,5;1,2], а в в кінці періоду – в діапазоні [0,4; 0,9] (рис. 4).

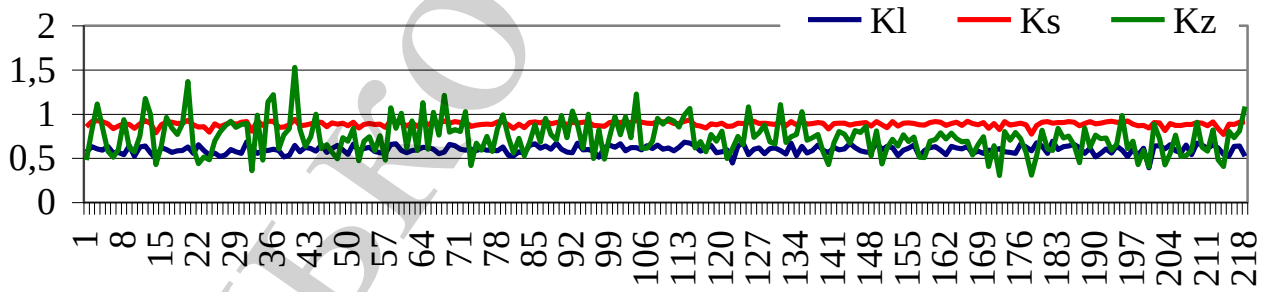


Рис. 4. Порівняння розподілу коефіцієнтів мовлення K_l, K_s та K_z

Аналогічно порівняємо розподіли індексів винятковості та концентрації (рис. 5). Якщо розмах розподілу суттєво не змінюється в часі для I_{wt}, то для I_{kt} є фіксовані значні зміни. З часом автор цих робіт все частіше повторюють деякі терміни в своїх роботах понад 10 разів, звужуючи коло своїх досліджень. На рис. 5, *г* поданий результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр. як мінімальне, максимальне та середнє значення за цей період (визначення коливання значень в цьому часовому проміжку). Більш суттєве коливання спостерігаємо за K_z (рис. 6).

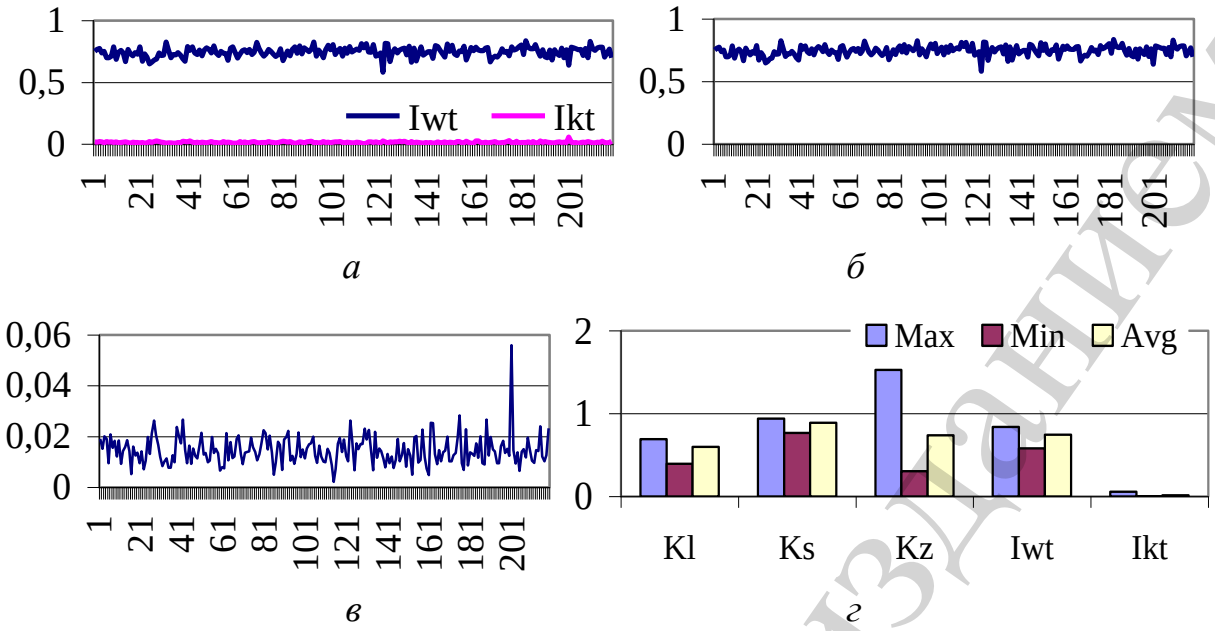


Рис. 5. Розподіл індексів мовлення для: *а* – обох індексів; *б* – I_{wt}; *в* – I_{kt}; *г* – мінімальне, максимальне та середнє значення для всіх коефіцієнтів

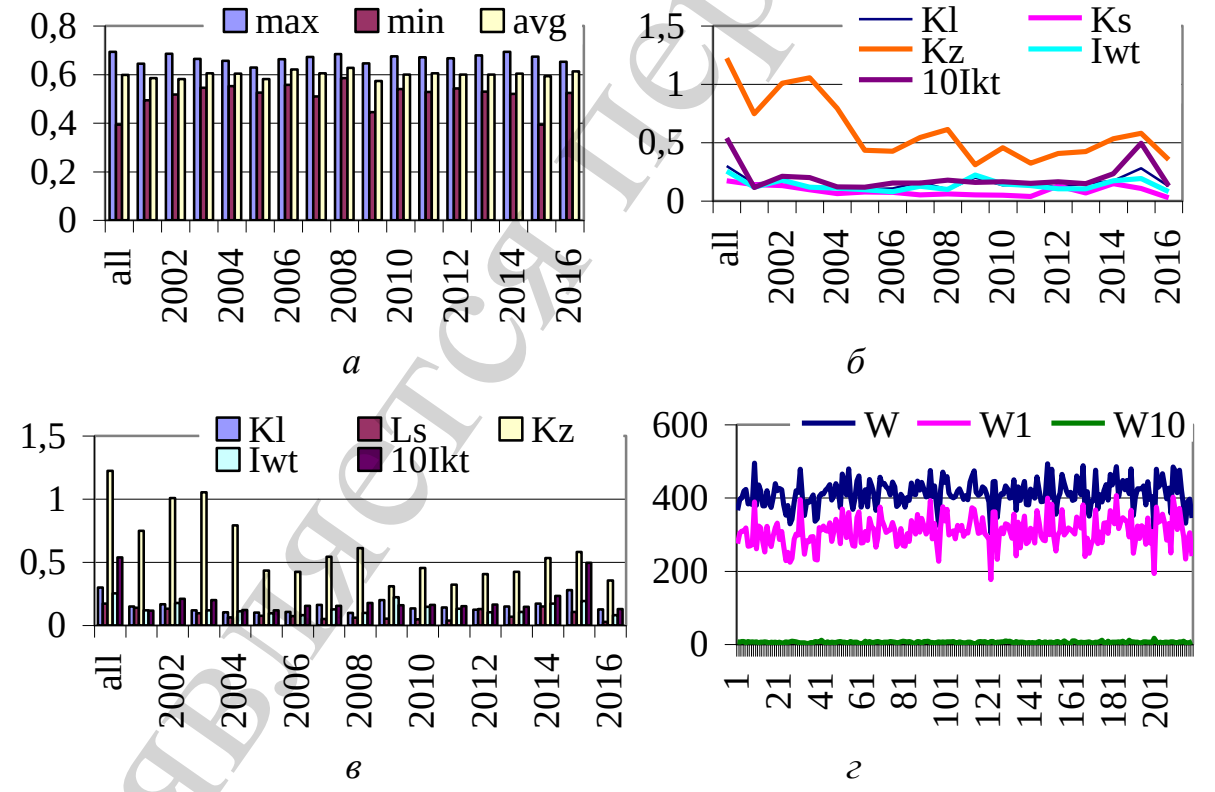


Рис. 6. Результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: *а* – мінімальне, максимальне та середнє значення за цей період для K_l; *б* – графік динаміки зміни коефіцієнтів за визначений період; *в* – гістограма динаміки зміни всі коефіцієнтів за визначений період; *г* – вживання словоформ (всіх, по 1 разу та понад 10 разів)

Окремо проаналізуємо розподіл використання всіх словоформ (рис. 6, *г*), слів по одному разу, слів понад 10 разів, вжитих в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр. (рис. 7, *а*). На рис. 7, *б* поданий аналіз вживання прийменників, сполучників та окремих речень в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр., де *Z* – кількість прийменників, *S* – кількість сполучників, *P* – кількість окремих речень. Згідно рис. 7, *в* з часом автори вживають коротші речення для опису предметної області, ніж на початках досліджуваного періоду. Якщо кількість прийменників зменшується, то розподіл вживання сполучників суттєво не зменшується (рис. 7, *д*). На рис. 8, *а–б* поданий аналіз зміни динаміки вживання слів в досліджуваних текстах за визначений період. На рис. 8, *в*–9, *г* поданий результат аналізу зміни динаміки вживання прийменників, сполучників та речень в досліджуваних текстах за визначений період.

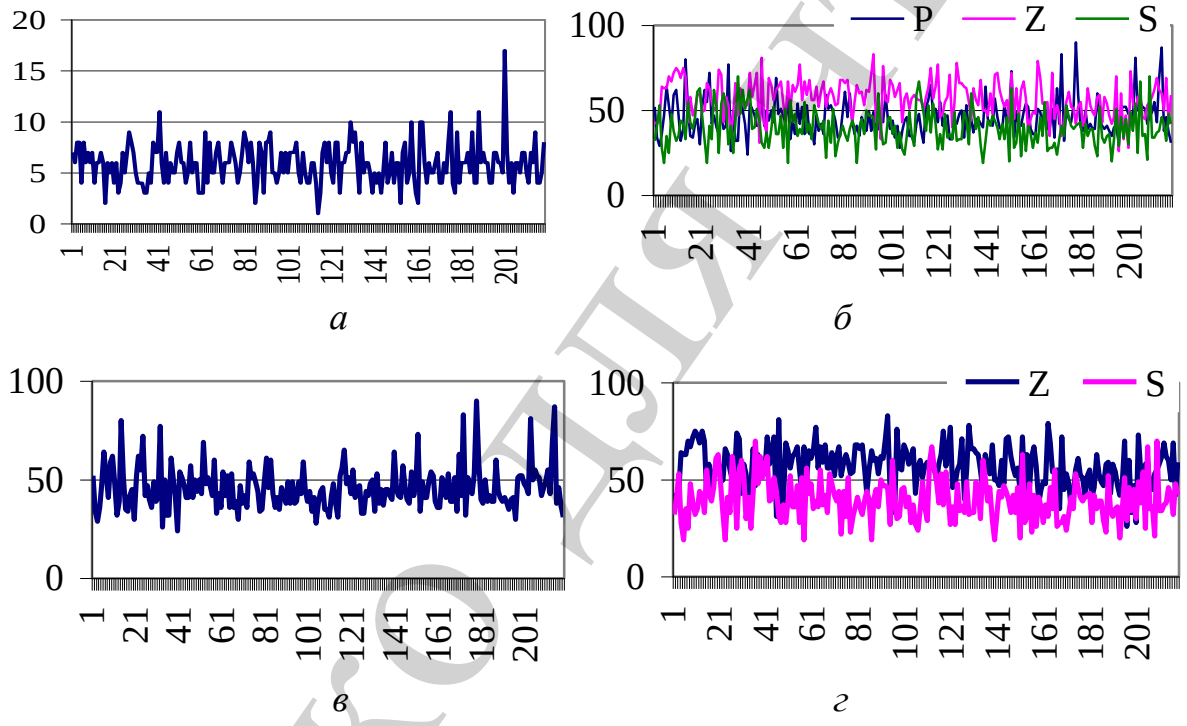


Рис. 7. Аналіз частоти вживання слів: *а* – понад 9 разів (W_{10}); *б* – параметрів зв'язності мовлення; *в* – речень; *г* – прийменників, та сполучників

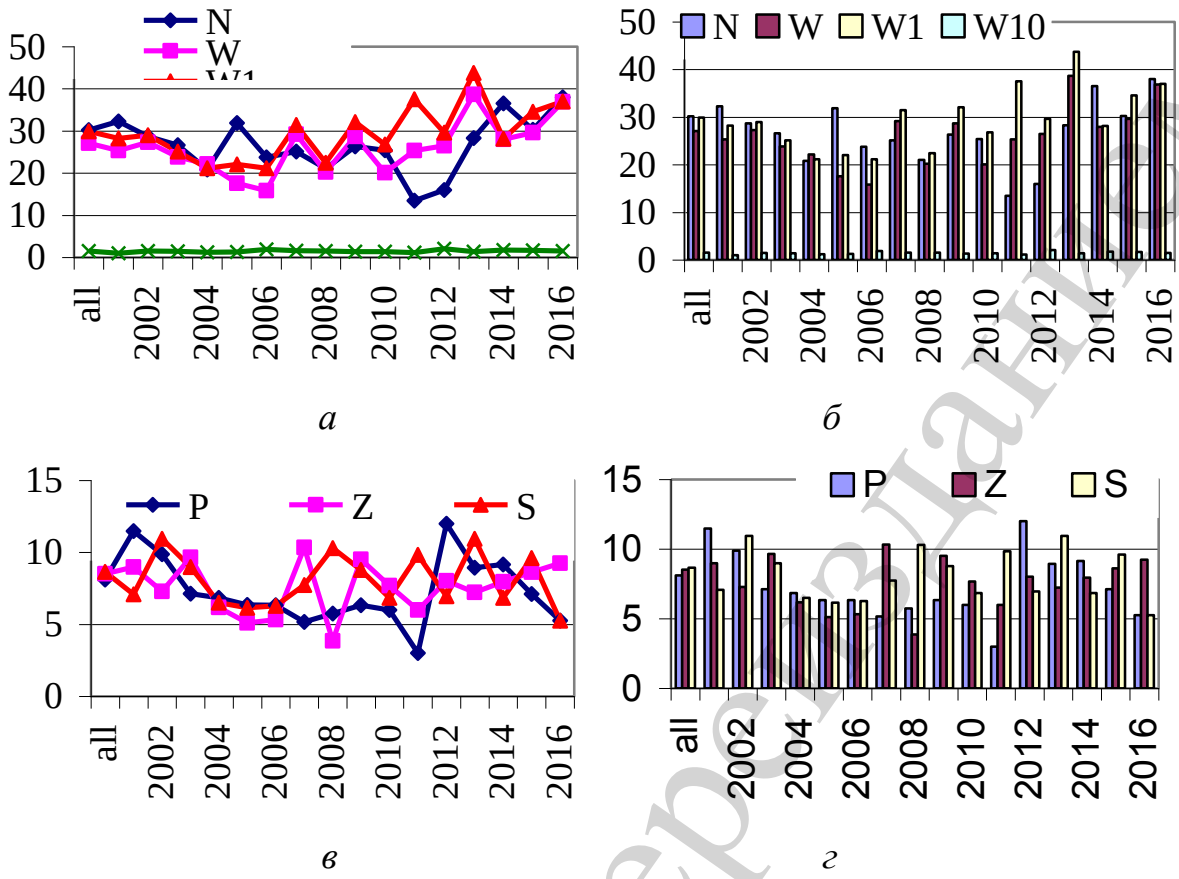


Рис. 8. Результат аналізу зміни динаміки вживання слів в досліджуваних текстах за визначений період: *а* – динаміка зміни параметрів мовлення в часі; *б* – розподіл значень параметрів мовлення за обумовлений період досліджень; *в* – динаміка зміни вживання сполучень, прийменників та речень в досліджуваних текстах; *г* – розподіл значень вживання сполучень, прийменників та речень за визначений період досліджень авторських стилів

Довели, що існує динаміка зміни не лише коефіцієнтів мовлення авторського тексту за визначений період його творчості. Також є динаміка зміни і окремих складових, як кількість вживання словоформ на загальну кількість слів, сполучників та прийменників, речень у визначеному обсязі уривку, словоформ, які вживані лише один раз, та які вживані понад 10 разів.

7. Обговорення результатів досліджень визначення автора в україномовних науково-технічних текстах

Для більш точного визначення величини приросту кожного із досліджуваного параметру необхідно провести більш суттєве дослідження на більшій вибірці як самих одноосібних творів, так збільшити діапазон дослідження творчості різних авторі на більший часовий проміжок творчості.

Далі проаналізуємо вибірку за авторським стилем та оберемо найкращий алгоритм для визначення стилю автора. На рис. 9, *а* графік відображає визначення стилю автора по коефіцієнтах мовлення. На рис. 9, *б* графік із накопиченням відображає зміни загальної суми за коефіцієнтами мовлення. На рис. 9,

в нормований графік відображає зміну вкладення кожного значення за коефіцієнтами мовлення.

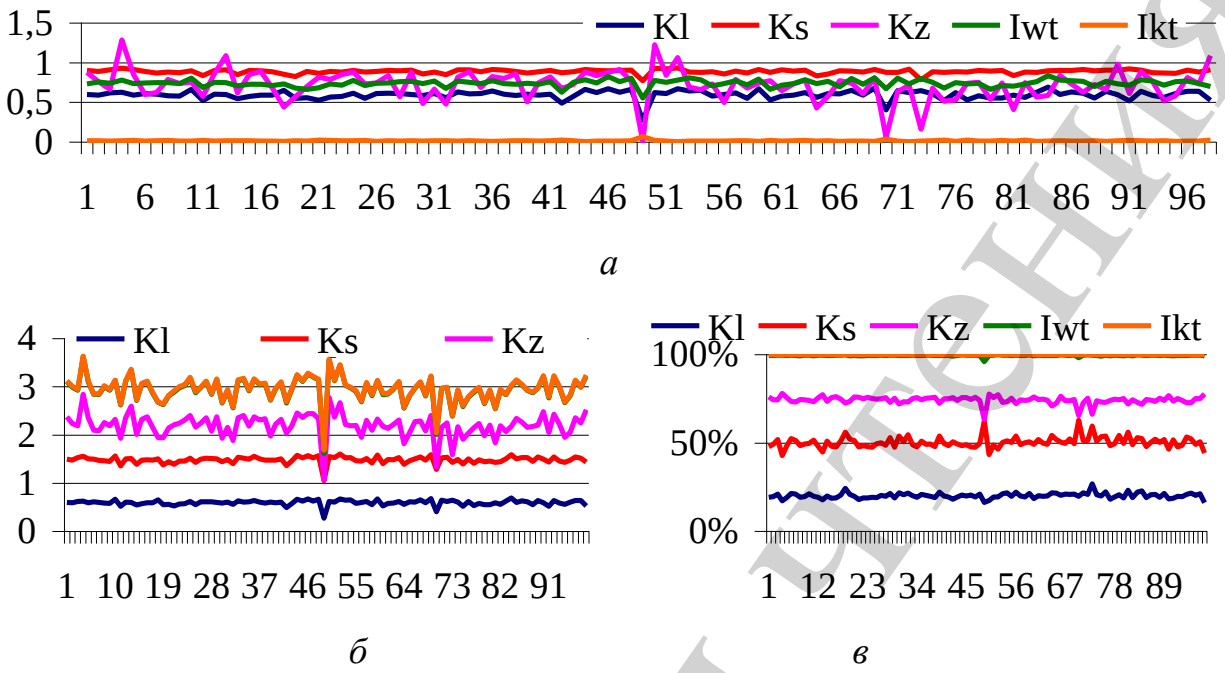


Рис. 9. Детальний аналіз: *а* – процесу у часі визначення стилю автора по коефіцієнтах мовлення; *б* – зміни загальної суми за коефіцієнтами мовлення; *в* – зміни вкладення кожного значення за коефіцієнтами мовлення

Як бачимо, коефіцієнти авторського мовлення окрім *Kz* значно не змінюються в залежності від стилю конктерного автора для україномовних науково-технічних текстів. Або зміни є в млих межах, що ускладнює процес ідентифікації особливостей стилю мовлення конкретного автора в множині аналізованих авторських стилів. І чим більшою буде така множина, тим складнішою буде процес ідентифікації стилю конкретного автора без додаткових параметрів аналізу. Тоді проаналізуємо вибірку за авторським стилем за додатковими параметрами як загальна кількіст речень в однакох за обсягом уривків, кількість слів у вибірці, частотність та пояка прийменників та сполучників. На рис. 10 графік відображає визначення стилю автора по додаткових параметрах авторського мовлення.

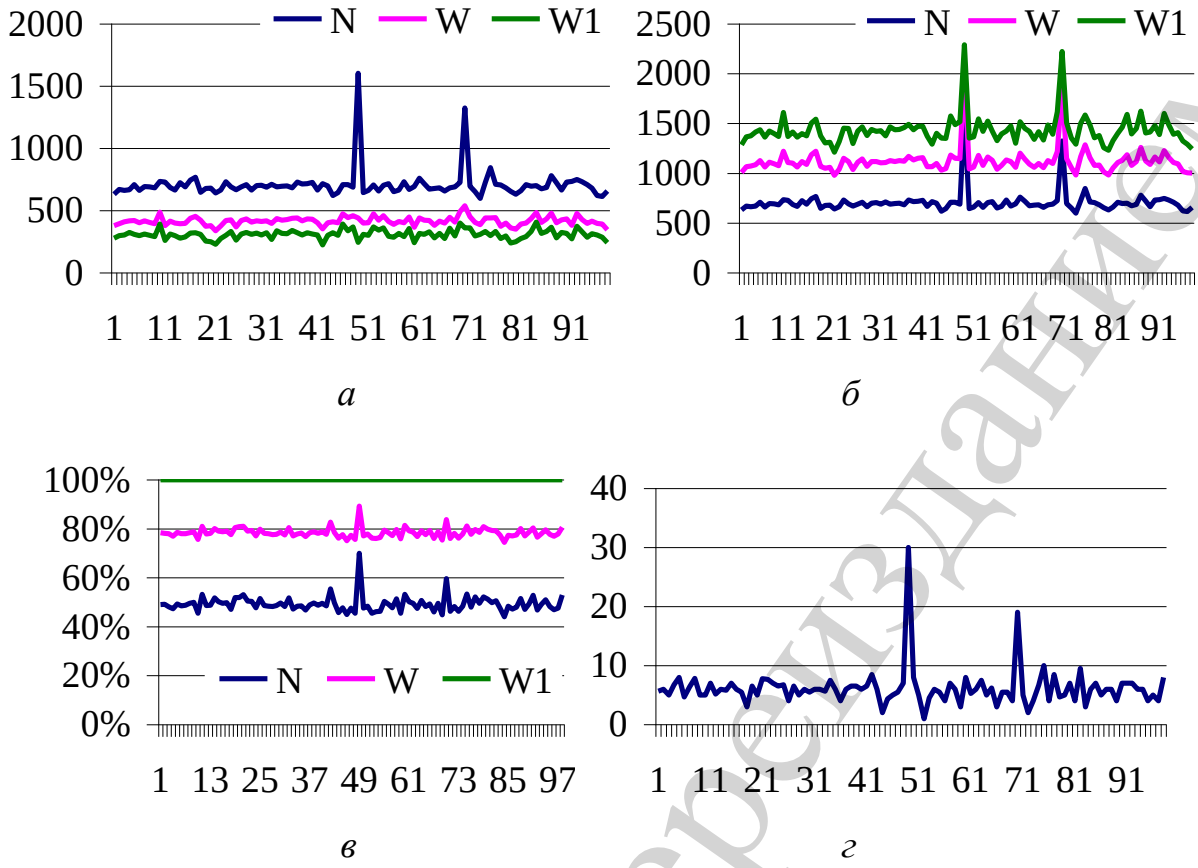
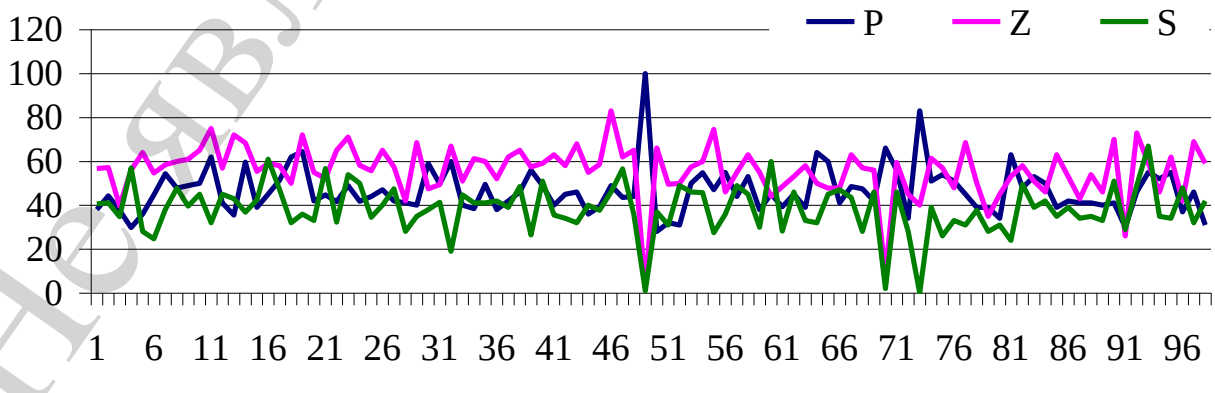
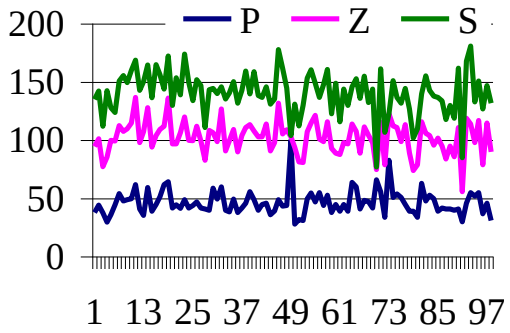


Рис. 10. Детальний аналіз: *а* – процесу визначення стилю автора по параметрах мовлення; *б* – зміни загальної суми за коефіцієнтами мовлення; *в* – зміни вкладення кожного значення за коефіцієнтами мовлення; *г* – зміни параметру як частота появи слова понад 10 разів (W10)

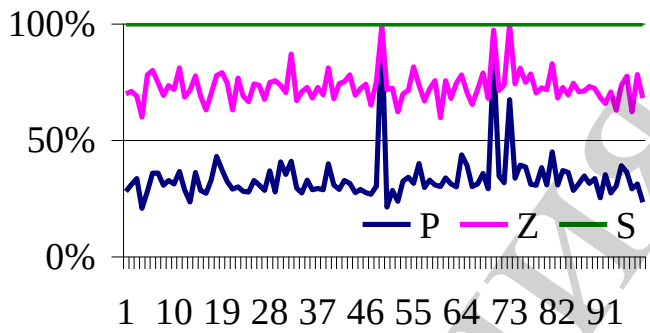
На рис. 10, *б* графік із накопиченням відображає зміни загальної суми за параметрами. На рис. 10, *в* нормований графік відображає зміну вкладення кожного значення за параметрами. Як бачимо введення додаткови параметрів зменшить множину авторів, стилі мовлення яких подібні для україномовного науково-технічного стилю публікацій. Введмо ще додаткові параметри як кількість речень, сполучників та прийменників (рис. 11) та проаналізуємо динаміку (табл. 5).



а



б



в

Рис. 11. Детальний аналіз: *а* – процесу визначення стилю автора по параметрах мовлення; *б* – зміни загальної суми за коефіцієнтами мовлення; *в* – зміни вкладення кожного значення за коефіцієнтами мовлення

Таблиця 5

Результат роботи алгоритму аналізу стилю автора публікації на інформаційному ресурсі Victana [25]

№	N	W	W_1	W_{10}	P	Z	S	K_l	K_s	K_z	I_{wt}	I_{kt}
1	671,3	395,6	299	6	44,2	57,1	41,1	0,59	0,89	0,76	0,76	0,015
2	662,5	410,3	303	5	37,8	39,8	34,8	0,61	0,9	0,67	0,74	0,012
3	668,8	418,3	325,8	6,8	29,8	56	57	0,63	0,93	1,28	0,78	0,016
4	708	419	309	8	36	64	28	0,59	0,91	0,85	0,74	0,019
5	661,1	402,7	299,7	4,7	44,7	54,7	24,8	0,61	0,89	0,6	0,74	0,012
6	694,5	417,4	313,1	6,4	54,3	58,5	38,1	0,6	0,87	0,62	0,75	0,015
7	691,8	403,4	301,6	7,8	47,8	60	47,8	0,58	0,88	0,79	0,75	0,019
8	682,5	394,2	291	5	49	61	39,7	0,58	0,88	0,74	0,74	0,013
9	733,5	486,5	392	5	50	65	45	0,66	0,9	0,76	0,8	0,01
10	729	380	261	7	62	75	32	0,52	0,84	0,58	0,69	0,018
11	686,5	414,5	312,6	5,9	41,1	56,9	45	0,6	0,9	0,86	0,75	0,012
12	665,5	399	299	6	35,5	72	43	0,6	0,91	1,09	0,75	0,015
13	724,2	394,2	278,8	5,8	59,6	68,4	36,8	0,55	0,85	0,61	0,71	0,015
14	691	396,7	289	7	39	55,3	42,3	0,57	0,9	0,85	0,73	0,018
15	745	439	319	6	45	59	61	0,59	0,9	0,89	0,73	0,014
16	768	452,5	323	5,5	51,5	58	47	0,59	0,89	0,68	0,71	0,012
17	647	422	308	3	62	50	32	0,65	0,85	0,44	0,73	0,007
18	677,5	373,5	255	6,5	64,5	72	36	0,55	0,86	0,57	0,68	0,018
19	680	379	251	5	42	55	33	0,56	0,89	0,7	0,66	0,013
20	642	337,5	230,3	7,8	44,8	52,3	56,8	0,52	0,87	0,81	0,68	0,023
21	665	376	275,7	7,7	41,7	65	32,3	0,57	0,89	0,79	0,73	0,02
22	731	420	301	7	49	71	54	0,57	0,88	0,85	0,72	0,017
23	691,7	425,7	331,3	6,5	41,8	58,2	50	0,62	0,9	0,88	0,78	0,015
24	668,8	368,3	262,5	6,8	44	55,8	34,5	0,55	0,88	0,73	0,71	0,018
25	691	421	311	4	47	65	40	0,6	0,89	0,74	0,74	0,01

26	708,5	434	323,5	6,5	42	57,5	47,5	0,61	0,9	0,84	0,75	0,015
27	665	406	309	5	41	42	28	0,61	0,9	0,57	0,76	0,012
28	700	418,5	320,5	6	40	68,5	35	0,6	0,9	0,88	0,77	0,014
29	704,5	412	303,5	5,5	59	47,5	38	0,58	0,86	0,49	0,74	0,013
30	688,8	416,8	321,9	6	49,7	49,3	41,3	0,6	0,88	0,67	0,77	0,016
31	711	396	268	6	60	67	19	0,56	0,85	0,48	0,68	0,015
32	691	436,7	336,7	5,7	40	51	44,7	0,63	0,91	0,82	0,77	0,013
33	695	422,5	318,3	7,5	38,5	61,3	41	0,6	0,91	0,89	0,75	0,018
34	699	427	314	6	49,5	60	41	0,61	0,88	0,69	0,74	0,014
35	683	438	339	4	38	52	42	0,64	0,91	0,82	0,77	0,009
36	730	440	323	6	42	62	39	0,6	0,9	0,8	0,73	0,014
37	714,5	418,5	304,5	6,5	46	65	48,5	0,59	0,89	0,86	0,73	0,016
38	717,5	433,5	321,5	6,5	56	57,5	26,5	0,6	0,87	0,5	0,74	0,015
39	728	430	313	6	49	59	51	0,59	0,89	0,75	0,73	0,014
40	666	401,5	305	6,5	40	63	35,5	0,6	0,9	0,82	0,76	0,016
41	715,5	352	223,5	8,5	45	58	34	0,49	0,87	0,68	0,63	0,024
42	699	401	302	6	46	68	32	0,57	0,89	0,72	0,75	0,015
43	620	411	323	2	36	55	40	0,66	0,91	0,88	0,79	0,005
44	645	403	302,3	4,3	39,3	58,7	37,7	0,62	0,9	0,84	0,74	0,011
45	708	475	392	5	49	83	46	0,67	0,9	0,88	0,83	0,011
46	708	442,5	336,5	5,5	43,5	62	56,5	0,63	0,9	0,91	0,76	0,012
47	689	458	369	7	44	65	36	0,66	0,9	0,77	0,81	0,015
48	1602	442	245	30	100	3	1	0,28	0,77	0,01	0,55	0,068
49	644	400	310	8	28	66	37	0,62	0,93	1,23	0,78	0,02
50	661,5	402,5	302	5	32	49,5	31	0,6	0,92	0,84	0,75	0,012
51	705	474	369	1	31	50	49	0,67	0,93	1,06	0,78	0,002
52	656	422,5	341,5	4,5	50	57,5	46	0,64	0,88	0,69	0,81	0,011
53	704,8	458,8	360	6	54,8	60	45,8	0,65	0,88	0,66	0,78	0,013
54	716	413,5	293	5,5	47	74,5	27,5	0,58	0,89	0,73	0,71	0,013
55	652	389	287	4	55	46	36	0,6	0,86	0,5	0,74	0,01
56	666	412	318	7	44	55	49	0,62	0,89	0,79	0,77	0,017
57	732	402	290	6	53	63	45	0,55	0,87	0,68	0,72	0,015
58	670	449	356	3	38	55	30	0,67	0,92	0,75	0,79	0,007
59	693	366	242	8	45	44	60	0,53	0,88	0,77	0,66	0,022
60	761	440	315,8	5,3	39,3	48,5	28,3	0,58	0,91	0,65	0,71	0,012
61	717	422	310	6	45	53	46	0,59	0,89	0,73	0,73	0,014
62	673,5	419	329	7,5	39	58	33	0,62	0,91	0,78	0,79	0,018
63	679	381	280	5	64	50	32	0,56	0,83	0,43	0,73	0,013
64	682,6	416,2	318	6,2	60	47,8	45	0,6	0,86	0,59	0,76	0,015
65	658	399	277	3	41	48	47	0,6	0,9	0,78	0,69	0,008
66	683	446	357	5,5	48,5	63	43,5	0,65	0,89	0,74	0,8	0,012
67	689,5	407,5	296	5,5	47,5	57	28	0,59	0,88	0,61	0,73	0,014
68	726	493	399	4	42	56	46	0,68	0,91	0,81	0,81	0,008

69	1325	538	360	19	66	9	2	0,4	0,88	0,06	0,67	0,035
70	697	450	361,5	5	56	59,5	46	0,65	0,88	0,63	0,8	0,011
71	652	405	296	2	34	45	28	0,62	0,92	0,72	0,73	0,005
72	598	386	309	4	83	40	0	0,65	0,78	0,16	0,8	0,01
73	726,3	441,3	332,3	6,7	51	61,3	39	0,6	0,88	0,68	0,75	0,015
74	846	440	299	10	54	57	26	0,52	0,88	0,51	0,68	0,023
75	712,5	442,5	331,5	4	51	48	33	0,62	0,88	0,53	0,75	0,009
76	706	374	275	8,5	45	68,5	31	0,53	0,88	0,74	0,73	0,023
77	682,3	398,7	296,3	4,7	39	50,3	37,7	0,58	0,9	0,75	0,74	0,012
78	654	361	240	5	39	35	28	0,55	0,89	0,54	0,66	0,014
79	631	350	249	7	34	45	31	0,55	0,9	0,75	0,71	0,02
80	661	391	275	4	63	53	24	0,59	0,84	0,41	0,7	0,01
81	709,5	399	292,5	9,5	48	58	49,5	0,56	0,88	0,75	0,73	0,024
82	695	436	332	3	53	51	39	0,63	0,88	0,57	0,76	0,007
83	700	485	406	6	50	46	42	0,69	0,9	0,59	0,84	0,012
84	674	404	316	7	39	63	35	0,9	0,9	0,84	0,78	0,017
85	685	432	333	5	42	53	39	0,63	0,9	0,73	0,77	0,012
86	780	479	366	6	41	43	34	0,61	0,91	0,63	0,76	0,013
87	723	401	280	6	41	54	35	0,55	0,9	0,72	0,7	0,015
88	665	425	324	4	40	46	33	0,64	0,91	0,66	0,76	0,009
89	730	433	317	7	41	70	51	0,59	0,91	0,98	0,73	0,016
90	734	381	273	7	30	26	29	0,52	0,92	0,61	0,72	0,018
91	749	478	375	7	46	73	49	0,64	0,9	0,88	0,78	0,015
92	732	429	329	6	55	59	67	0,59	0,87	0,76	0,77	0,014
93	709	398	285	6	52	46	35	0,56	0,87	0,52	0,72	0,015
94	680	414	314	4	55	62	34	0,6	0,87	0,58	0,76	0,01
95	622	397	305	5	37	42	48	0,64	0,91	0,81	0,77	0,013
96	614	391	287	4	46	69	32	0,64	0,88	0,73	0,73	0,01
97	658	345	241	8	31	59	42	0,52	0,91	1,07	0,7	0,023
98	631,3	377,7	277,7	5,7	38	56,7	40,7	0,6	0,9	0,88	0,73	0,015

В табл. 5 наведені результати аналізу стилю 94 авторів на одноосібних працях (понад 200 одноосібних робіт) технічного спрямування за період 2001–2017 рр. Для кожного автора виведено середньоарифметичне значення кожного коефіцієнта та параметра мовлення на основі аналізу декілької його робіт за цей визначений період. Також проаналізовані стилі 4-х статтей одного авторського колективу під № 95–98 (в таблиці виділено жовтим кольором), частина авторів яких є в табл. 5 під № 6 та 30 (в таблиці виділено синім кольором).

Однак замала вибірка текстів для аналізу (понад 200) та кількості авторів (94) не гарантує точних результатів. Дослідження має бути продовжене на більшій кількості текстів, до яких незавжди маємо доступ. В подальшому необхідно також вдосконалити метод зарахунок аналізу текстів методами стилем атрії та глотохронології.

6. Висновки

1. Розроблено метод визначення автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту на основі аналізу кожного слова з врахуванням його частини мови та відмінювання. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частни мови. Для цього прводився аналіз флекцій цих слів для класифікації, виділення основи для формування відповідних алфавітно–частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації. Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Його особливостями є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Наведено теоретичне та експериментальне обгрунтування методу контент–моніторінгу та визначення стопових слів україномовного тексту. Метод спрямовано на автоматичне виявлення значущих стопових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації парсингу текстового контенту науково-технічного спрямування.

2. Запропоновано підхід до розроблення програмного забезпечення контент-моніторінгу для визначення автора в україномовних науко–технічних текстах на основі NLP, стилементії та Web Mining. Проаналізовано розробленою системою понад 200 одноосібних наукових публікацій зі всіх номерів Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (Україна) за період 2001–2017 рр. Досліджено внутрішню «динаміку» цих текстів довільно обраних авторів через аналіз коефіцієнтів зв'язності мовлення, лексичної різноманітності та синтаксичної складності, а також індексів концентрації та винятковості для перших k, n та m (без заголовка) слів авторського уривку та аналізованого.

3. Досліджено результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. На основі розробленого програмного забезпечення отримано результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення та аналізу стопових слів в україномовних наукових текстах технічного профілю на основі технології Web Mining. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури. Подальшого експериментального дослідження

потребує апробація запропонованого методу для визначення стилю автора з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

Література

1. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology / Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4, Issue 2 (88). P. 10–19. doi: https://doi.org/10.15587/1729-4061.2017.107512
2. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining / Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2, Issue 2 (86). P. 14–23. doi: https://doi.org/10.15587/1729-4061.2017.98750
3. The method of formation of the status of personality understanding based on the content analysis / Lytvyn V., Pukach P., Bobyk I., Vysotska V. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5, Issue 2 (83). P. 4–12. doi: https://doi.org/10.15587/1729-4061.2016.77174
4. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach / Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3, Issue 2 (87). P. 11–17. doi: https://doi.org/10.15587/1729-4061.2017.103630
5. Analysis of statistical methods for stable combinations determination of keywords identification / Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. // Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2, Issue 2 (92). P. 23–37. doi: https://doi.org/10.15587/1729-4061.2018.126009
6. Khomytska I., Teslyuk V. 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). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589887
7. Khomytska I., Teslyuk V. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level // Advances in Intelligent Systems and Computing. 2016. P. 149–163. doi: https://doi.org/10.1007/978-3-319-45991-2_10
8. Mobasher B. Data Mining for Web Personalization // Lecture Notes in Computer Science. 2007. P. 90–135. doi: https://doi.org/10.1007/978-3-540-72079-9_3
9. Dinucă C. E., Ciobanu D. Web Content Mining // Annals of the University of Petroşani. Economics. 2012. Vol. 12, Issue 1. P. 85–92.
10. Xu G., Zhang Y., Li L. Web Content Mining // Web Mining and Social Networking. 2010. P. 71–87. doi: https://doi.org/10.1007/978-1-4419-7735-9_4

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

12. Анисимов А., Марченко А. Система обработки текстов на естественном языке // Искусственный интеллект. 2002. № 4. С. 157–163.

13. Перебийніс В. Математична лінгвістика. Українська мова. Київ, 2000. С. 287–302.

14. Бук С. Основи статистичної лінґвістики. Львів, 2008. 124 с.

15. Перебийніс В. Статистичні методи для лінгвістів. Вінниця, 2013. 176 с.

16. Браславский П. И. Интеллектуальные информационные системы. URL: http://www.kansas.ru/ai2006/

17. Ланде Д., Жигало В. Підхід до рішення проблем пошуку двомовного плагіату // Проблеми інформатизації та управління. 2008. № 2 (24). С. 125–129.

18. Варфоломеев А. Психосемантика слова и лингвостатистика текста. Калининград, 2000. 37 с.

19. Сушко С., Фомичова Л., Барсуков Є. Частоти повторюваності букв і біграм у відкритих текстах українською мовою // Захист інформації. 2010. Т. 12, № 3 (48). doi: https://doi.org/10.18372/2410-7840.12.1968

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

21. Кочерган М. Вступ до мовознавства. Київ, 2005. 368 с.

22. Родионова Е. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика. 2008. № 7. С. 118–127.

23. Мещеряков Р. В., Васюков Н. С. Модели определения авторства текста. URL: http://db.biysk.secna.ru/conference/conference.conference.doc_download?id_thesis_dl=427

24. Морозов Н. А. Лингвистические спектры. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3

25. Victana. URL: http://victana.lviv.ua/nlp/linhvometriia

26. Method of Integration and Content Management of the Information Resources Network / Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. // Advances in Intelligent Systems and Computing. 2017. P. 204–216. doi: https://doi.org/10.1007/978-3-319-70581-1_14

27. Information resources processing using linguistic analysis of textual content / Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. // 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). 2017. doi: https://doi.org/10.1109/idaacs.2017.8095038

28. The risk management modelling in multi project environment / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098730

29. Peculiarities of content forming and analysis in internet newspaper covering music news / Korobchinsky M., Chyrun L., Chyrun L., Vysotska V. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098735

30. Intellectual system design for content formation / Naum O., Chyrun L., Vysotska V., Kanishcheva O. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098753

31. The Contextual Search Method Based on Domain Thesaurus / Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. // Advances in Intelligent Systems and Computing. 2017. P. 310–319. doi: https://doi.org/10.1007/978-3-319-70581-1_22

32. Марченко О. Моделювання семантичного контексту при аналізі текстів на природній мові // Вісник Київського університету. 2006. № 3. С. 230–235.

33. Jivani A. G. A Comparative Study of Stemming Algorithms // Int. J. Comp. Tech. Appl. 2011. Vol. 2, Issue 6. P. 1930–1938.

34. Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis / Mishler A., Crabb E. S., Paletz S., Hefright B., Golonka E. // Communications in Computer and Information Science. 2015. P. 639–644. doi: https://doi.org/10.1007/978-3-319-21380-4_108

35. Родионова Е. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика. 2008. № 7. С. 118–127.

36. Бублейник Л. Особливості художнього мовлення. Луцьк, 2000. 179 с.

37. Kowalska K., Cai D., Wade S. Sentiment Analysis of Polish Texts // International Journal of Computer and Communication Engineering. 2012. P. 39–42. doi: https://doi.org/10.7763/ijcce.2012.v1.12

38. Kotsyba N. The current state of work on the Polish–Ukrainian Parallel Corpus (PolUKR) // Organization and Development of Digital Lexical Resources. 2009. P. 55–60.

39. Single-frame image super-resolution based on singular square matrix operator / Rashkevych Y., Peleshko D., Vynokurova O., Izonin I., Lotoshynska N. // 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). 2017. doi: https://doi.org/10.1109/ukrcon.2017.8100390

40. Learning-Based Image Scaling Using Neural-Like Structure of Geometric Transformation Paradigm / Tkachenko R., Tkachenko P., Izonin I., Tsymbal Y. // Studies in Computational Intelligence. 2017. P. 537–565. doi: https://doi.org/10.1007/978-3-319-63754-9_25

41. Vysotska V. Linguistic analysis of textual commercial content for information resources processing // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). 2016. doi: https://doi.org/10.1109/tcset.2016.7452160

42. Detection of near dublicates in tables based on the locality-sensitive hashing method and the nearest neighbor method / Lizunov P., Biloshchytskyi A., Kuchansky A., Biloshchytska S., Chala L. // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6, Issue 4 (84). P. 4–10. doi: https://doi.org/10.15587/1729-4061.2016.86243

43. Conceptual model of automatic system of near duplicates detection in electronic documents / Biloshchytskyi A., Kuchansky A., Biloshchytska S., Dubnytska A. // 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). 2017. doi: https://doi.org/10.1109/cadsm.2017.7916155

44. Vysotska V., Rishnyak I., Chyryn L. Analysis and Evaluation of Risks in Electronic Commerce // 2007 9th International Conference – The Experience of Designing and Applications of CAD Systems in Microelectronics. 2007. doi: https://doi.org/10.1109/cadsm.2007.4297570

45. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589909

46. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589902

47. Content linguistic analysis methods for textual documents classification / Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589903

48. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325446

49. Vysotska V., Chyrun L. Analysis features of information resources processing // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). 2015. doi: https://doi.org/10.1109/stc-csit.2015.7325448

50. Application of sentence parsing for determining keywords in Ukrainian texts / Vasyl L., Victoria V., Dmytro D., Roman H., Zoriana R. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098797

51. Maksymiv O., Rak T., Peleshko D. Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency // International Journal of Intelligent Systems and Applications. 2017. Vol. 9, Issue 2. P. 42–48. doi: https://doi.org/10.5815/ijisa.2017.02.06

52. Peleshko D., Rak T., Izonin I. Image Superresolution via Divergence Matrix and Automatic Detection of Crossover // International Journal of Intelligent

Systems and Applications. 2016. Vol. 8, Issue 12. P. 1–8. doi: https://doi.org/10.5815/ijisa.2016.12.01

53. The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing / Bazylyk O., Taradaha P., Nadobko O., Chyrun L., Shestakevych T. // 2012 11th International Conference on "Modern Problems of Radio Engineering, Telecommunications and Computer Science" (TCSET). 2012. P. 107–108.

54. The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture / Bondariev A., Kiselychnyk M., Nadobko O., Nedostup L., Chyrun L., Shestakevych T. // TCSET'2012. 2012. P. 159.

55. Riznyk V. Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 129–148. doi: https://doi.org/10.1007/978-3-319-45991-2_9

56. Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System / Teslyuk V., Beregovskyi V., Denysyuk P., Teslyuk T., Lozynskyi A. // International Journal of Intelligent Systems and Applications. 2018. Vol. 10, Issue 1. P. 1–8. doi: https://doi.org/10.5815/ijisa.2018.01.01

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

58. Pasichnyk V., Shestakevych T. The model of data analysis of the psychophysiological survey results // Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 271–281. doi: https://doi.org/10.1007/978-3-319-45991-2_18

59. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 656–667. doi: https://doi.org/10.1007/978-3-319-70581-1_45

60. Chernukha O., Bilushchak Y. Mathematical modeling of random concentration field and its second moments in a semispace with erlangian disrtibution of layered inclusions // Task Quarterly. 2016. Vol. 20, Issue 3. P. 295–334.

61. Davydov M., Lozynska O. 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). 2017. doi: https://doi.org/10.1109/stc-csit.2017.8098734

62. Davydov M., Lozynska O. Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 89–100. doi: https://doi.org/10.1007/978-3-319-70581-1_7

63. Davydov M., Lozynska O. Linguistic models of assistive computer technologies for cognition and communication // 2016 XIth International Scientific

and Technical Conference Computer Sciences and Information Technologies (CSIT). 2016. doi: https://doi.org/10.1109/stc-csit.2016.7589898

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

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

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

67. Mykich K., Burov Y. 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

68. Kravets P. The control agent with fuzzy logic // Perspective Technologies and Methods in MEMS Design, MEMSTECH'2010 – Proceedings of the 6th International Conference. Lviv, 2010. P. 40–41.

69. On the Asymptotic Methods of the Mathematical Models of Strongly Nonlinear Physical Systems / Pukach P., Il'kiv V., Nytrebych Z., Vovk M., Pukach P. // Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 421–433. doi: https://doi.org/10.1007/978-3-319-70581-1_30

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

71. Kravets P. Game Model of Dragonfly Animat Self-Learning // Perspective Technologies and Methods in MEMS Design. 2016. P. 195–201.