ОГЛЯДИ

УДК 811.161.2'32'374
DOI: https://doi.org/10.17721/1728-2659.2024.36.21

Оксана ЗУБАНЬ, канд. філол. наук, доц.
ORCID ID: 0000-0002-2644-3892
e-mail: oxana.zuban@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

ФУНКЦІОНАЛ АВТОМАТИЧНОЇ СИСТЕМИ ЛІНГВОСТАТИСТИЧНОЇ АНАЛІТИКИ УКРАЇНСЬКОМОВНИХ МЕДІЙНИХ ТЕКСТІВ – TEXTATTRIBUTOR 1.0 (ІНСТРУКЦІЇ КОРИСТУВАЧЕВІ)

Представлено інструкції для користувача, які знайомлять із функціоналом і роботою автоматичної системи параметризації українськомовного медійного тексту TextAttributor 1.0. Ця система реалізована як вебзастосунок (http://ta.mova.info/), що дозволяє користувачеві в інтерактивному режимі здійснити лінгвостатистичний аналіз уведеного тексту й отримати статистичні дані про параметризацію українськомовного медійного тексту за 18-ма статистичними параметрами. Також функціонал системи генерує експертний висновок лінгвостатистичного аналізу тексту та графічне унаочнення стилеметричного порівняння одного або двох текстів із еталонними статистичними характеристиками медійного стилю української мови. Окремими модулями системи є: 1) модуль "Порівняння атрибуції текстів", у якому визначається ступінь схожості двох, обраних користувачем, текстів у завданні встановлення авторства; 2) модуль "Лінгвістична експертиза токсичності тексту", у якому користувач отримує систематизовані лінгвістичні та статистичні дані про токсичність українськомовного медійного тексту. Система розрахована на науковців і пересічних користувачів, яких цікавить аналітика текстової інформації з метою оцінки медіатекстів у розв'язанні таких завдань, як верифікація авторства, психолінгвометричне профілювання, моделювання стилів, фільтрування текстової інформації в автоматичну моніторингу інтернет-простору, відстежування поширювачів токсичних текстів. Вільний доступ до вебзастосунку TextAttributor 1.0, зручний інтерфейс і систематизація лінгвістичної експертної аналітики українськомовних медійних текстів відкривають широкі можливості користувачам для отримання необхідної інформації. Інструкцію для роботи у системі було оприлюднено на сайті вебзастосунку (квітень 2020 р.), але друкується ця інструкція вперше.

Ключові слова: ***український медійний текст, автоматична система лінгвістичного аналізу, стилеметрія, атрибуція тексту, лінгвостатистичний параметр, лінгвістична експертиза, токсичний текст.***

Вступ

Квантитативна лінгвістика як у системі філологічних дисциплін, так і в наукових дослідженнях українського мовознавства, зокрема і прикладної лінгвістики, не набула належної наукової значущості. Сучасні лінгвістичні та літературознавчі дослідження, у кращому випадку, переважно використовують лише кількісні підрахунки лінгвістичних явищ та одиниць, а дослідники оперують поняттями більше / менше з опорою на абсолютну частоту або її відсотковий еквівалент. Такий рівень "квантитативності" філологічних досліджень пояснюється багатьма факторами: відсутністю культури використання статистичних методів у філології, попри вагомі здобутки структурно-математичної лінгвістики в українському мовознавстві (Статистичні…, 1967), (Перебийніс, Муравицька, & Дарчук, 1985), (Бук, 2021), (Зубань, 2020), (Darchuk, Zuban, & Sorokin, 2024); ігноруванням кількісних методів дослідження порівняно із якісними; відсутністю базових дисциплін із лінгвостатистики у підготовці філологів у ЗВО України; трудомісткістю проведення лінгвостатистичного експерименту "вручну".

Із розвитком інформаційних технологій, зокрема комп'ютерної лінгвістики, здобутки якої уможливлюють проведення автоматичної параметризації тексту на всіх рівнях його організації, у сучасній філологічній науці відбулася зміна акцентів і статистичні методи починають частіше застосовуватися в дослідженнях. Тому, керуючись актуальними завданнями сучасної квантитативної лінгвістики й автоматичного опрацювання природної мови, колектив комп'ютерних лінгвістів, сформований на базі викладачів кафедри української мови та прикладної лінгвістики Київського національного університету імені Тараса Шевченка (Н. Дарчук, О. Зубань, В. Робейко, Ю. Цигвінцева, В. Сорокін, М. Сажок, М. Костіков), у межах проєкту "Визначення авторства анонімних українськомовних текстів із використанням методів штучного інтелекту в мережі Інтернет" (за фінансової підтримки Уряду Великої Британії) поставив мету – створити систему автоматичної статистичної параметризації українськомовного медійного тексту. Унаслідок плідної праці було створено систему автоматичного лінгвостатистичного аналізу, реалізовану у формі вебзастосунку TextAttributor 1.0 (TextAttributor 1.0, 2024), який був представлений широкому загалу науковців і користувачів 29.03.2024 р. в Укрінформі (Визначення…, 2024).

Система TextAttributor 1.0 працює у чотирьох завданнях: 1) статистичної параметризації тексту – базові лінгвостатистичні дані для виконання наступних завдань; 2) стилеметрії: визначення лінгвостатистичних ознак ідіостилю тексту; 3) атрибуції авторства: визначення ступеня схожості двох текстів; 4) сентимент-аналізу: визначення в тексті лексики з негативним сентиментом. До того ж, у межах системи за результатами виконання другого й четвертого завдання автоматично генеруються два висновки лінгвістичної експертизи. Поліфункційність системи TextAttributor 1.0 передбачає ознайомлення користувачів із принципами її роботи. Із цією метою було написано "Інструкцію користувачеві", яка розміщена на сторінці вебзастосунку (TextAttributor:..., 2024). Мета цієї оглядової статті ознайомити філологічну спільноту з цією інструкцією для розуміння функцій і аналітичних можливостей системи TextAttributor 1.0.

Результати

Яке призначення системи TextAttributor? Система виконує функцію інтерактивного аналітичного інструмента для проведення лінгвостатистичного аналізу з метою визначення атрибуції українськомовного медійного тексту в завданнях стилеметрії та встановлення авторства, а також із метою генерації висновку лінгвістичної експертизи тексту за статистичними параметрами атрибуції і негативною тональністю – токсичністю тексту.

ISSN 1728-2659 (Print), ISSN 2709-8494 (Online)

Система працює в режимі стилеметричного порівняння текстів лише на матеріалі текстів медійного стилю, проте лінгвостатистичний аналіз, без визначення індивідуальних ознак стилю автора, можна проводити на матеріалі українськомовного тексту будь-якого стилю.

На кого розрахована система TextAttributor? Система орієнтована на широке коло користувачів. Насамперед вона корисна науковцям: лінгвістам, літературознавцям, письменникам, журналістам, редакторам, видавцям, політологам, історикам, криміналістам, соціологам, психологам і представникам інших галузей науки – у завданні отримання об'єктивних даних про організацію тексту, порівняння текстів на основі цих точних даних із метою перевірки гіпотез різних наукових галузей, установлення авторства тексту, проведення стилеметричних досліджень. Також у сучасному просторі інтернет-комунікації ці завдання виходять за межі наукових досліджень і становлять зацікавленість пересічних користувачів у сфері аналітики отримуваної текстової інформації з метою оцінки медіатекстів та інформаційного захисту. Вільний доступ до сайту TextAttributor, зручний інтерфейс і систематизація аналітики відкривають широкі можливості користувачам для отримання необхідної інформації.

Інструкція користувачеві

Як увести текстові дані в системі TextAttributor?

1. Реєстрація у системі TextAttributor здійснюється за електронною поштою користувача, що створює індивідуальний акаунт, тому результати роботи в системі доступні лише одному користувачеві.

2. Уведення тексту для аналізу – лінк "Мої тексти" (рис. 1).

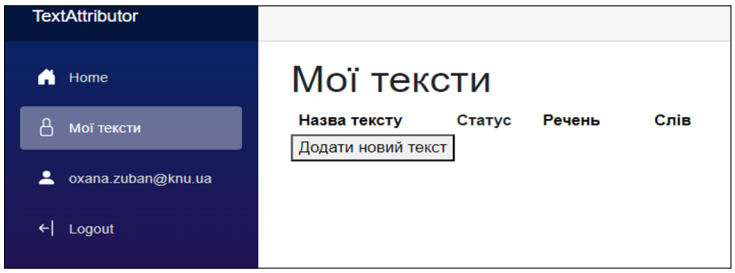


Рис. 1. Функціонал режиму "Мої тексти"

Після активації лінку "Додати новий текст" користувач переходить на сторінку для введення тексту в систему (рис. 2). Користувач копіює свій текст і вставляє назву тексту та сам текст у відповідні поля.

Рекомендація: зверніть увагу, що для отримання вірогідних статистичних даних аналізу за потреби варто здійснити технічне редагування тексту (розставити пробіли між словами, розділові знаки між реченнями, виправити одруківки тощо). Редагування можна здійснити до введення тексту в поле "Текст" або в самому полі.

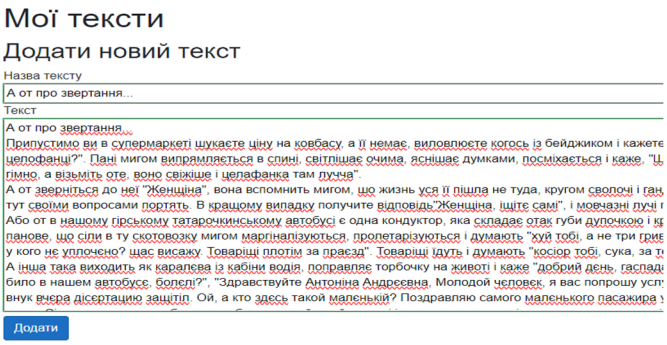


Рис. 2. Режим уведення тексту до системи

3. Оброблення тексту. Після активації команди "Додати" (рис. 2) починається автоматичне оброблення тексту, воно триває від кількох секунд до 5–10 хв залежно від обсягу завантаженого тексту. Інформація про завершення процесу автоматичного аналізу "Оброблено" з'являється в колонці "Статус" (рис. 1, 3), а в наступних колонках – інформація про кількість слів і речень. Активуючи лінк "Додати новий текст", користувач за описаною процедурою вводить нові тексти, аналіз яких доступний лише за акаунтом одного користувача. Сформований каталог текстів можна редагувати, видаляючи та завантажуючи нові тексти.

Рекомендація: зверніть увагу, якщо в колонці "Статус" залишається довго режим "Обробка" і не з'являється результат "Оброблено", потрібно оновити сторінку.

Яку інформацію отримує користувач за результатами автоматичного лінгвостатистичного аналізу? Після активації команди "Оброблено" (рис. 3) користувач потрапляє на сторінку результатів автоматичного оброблення того тексту, у рядку якого міститься активована команда. Ця сторінка сайту структурована за такими рубриками:

Індекси атрибуції тексту
Експертний висновок атрибуції тексту
Порівняння атрибуції текстів
Лінгвістична експертиза токсичності тексту
Висновки нейронної мережі

Як інтерпретувати індекси атрибуції тексту? Рубрика "Індекси атрибуції тексту" систематизує за табличним принципом обчислені статистичні параметри за введеним користувачем текстом (рис. 4). У першій колонці подається назва статистичного параметра, а в другій – назва введеного тексту та числові значення автоматично обчислених індексів. Система обчислює 18 статистичних

параметрів, із яких перші три є кількісними даними про обсяг словника, тексту та кількість речень у тексті. Також окремим рядком подано кількість слів аналізованого тексту, які не оброблені системою. Ці слова не входять до числового значення обсягу слів тексту і можуть свідчити про похибку лінгвостатистичного дослідження. До таких слів належать діалектизми, русизми, оказіоналізми та ін. Наприклад, на рис. 4 необроблених слів усього 4. Необроблені слова, а також деякі символи виділяються по тексту червоним кольором при активації команди – "Показати текст" у рубриці Лінгвістична експертиза токсичності тексту.

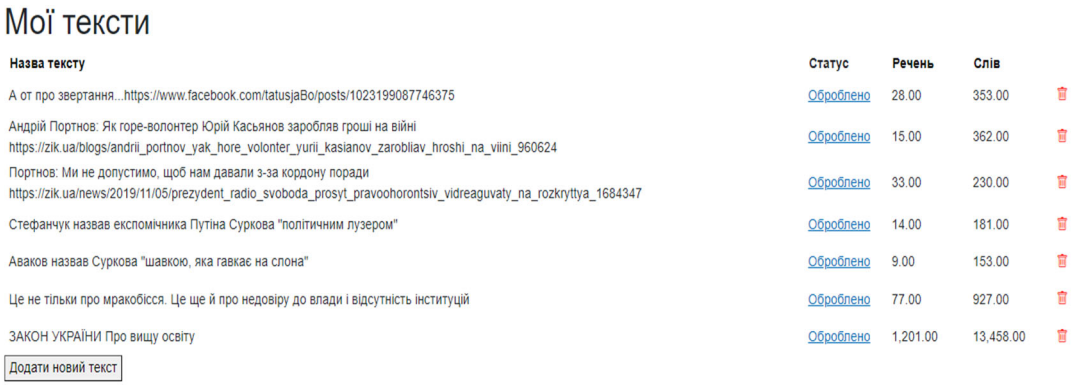


Рис. 3. Систематизація текстів користувача в каталог

Рис. 4. Фрагмент рубрики "Індекси атрибуції тексту"

Наступні 15 параметрів – статистичні індекси, згруповані так:

1) за формальними та граматичними лінгвістичними ознаками обчислювальних одиниць:

- кількісне співвідношення слів реєстру словника та обсягу текстових слововживань;
- кількісне співвідношення лексико-граматичних класів слів (частин мови) у тексті;
- кількісне співвідношення словосполучень та речень у тексті;

2) за психолінгвістичними ознаками;

3) за семантичною ознакою – негативним сентиментом тексту.

Рекомендація: зверніть увагу, що користувач може отримати в колонці 2 вірогідні дані про статистичні параметри тексту не лише для текстів медійного стилю, а й українськомовного тексту будь-якого стилю.

Біля кожного індексу у верхньому регістрі (рис. 4) є інтерактивна мітка "*i*" (інформаційна довідка). З активацією "*i*" в окремому вікні подається інформаційна довідка про лінгвостатистичну характеристику індексу та формулу його обчислення (рис. 5).

У третій колонці "Референс" (рис. 4) подано зіставлення отриманого числового значення індексу з еталонними (середніми) числовими значеннями медійного стилю. Це зіставлення унаочнено на шкалі, яка графічно відображає інтервал коливання середнього числового значення індексу в медійному стилі української мови (рис. 4): на шкалі відкладено нижнє і верхнє порогові числові значення, які формують відрізок довірчого інтервалу медійного стилю. Місце обчисленого (емпіричного) числового значення індексу позначено на шкалі синім перевернутим трикутником. У межах довірчого інтервалу містяться типові числові значення для медійних текстів, а вихід за межі цього інтервалу свідчить про те, що в аналізованому тексті нетипове для медійного стилю числове значення індексу. Тобто, якщо числове значення (синій трикутник) у межах інтервалу, то статистична характеристика за індексом підкоряється статистичним законам медійного стилю, тобто між еталоном медійного стилю й досліджуваним текстом немає істотних розходжень у частоті досліджуваного явища (нульова гіпотеза приймається), а якщо числове значення (синій трикутник) за межами інтервалу, то статистична характеристика

досліджуваного явища, представлена тим чи іншим індексом, свідчить про індивідуальні ознаки авторського стилю (нульова гіпотеза відхиляється). Потрібно також зауважити, що в системі TextAttributor довірчий інтервал для індексів медійного стилю було обчислено окремо для коротких текстів обсягом до 1000 слововживань (інформацію про обсяг тексту подано в попередніх рубриках, рис. 4, 7) та окремо для довгих текстів обсягом більше ніж 1000 слововживань. Це важливо для вірогідності статистичних даних, тому що залежно від обсягу тексту лінгвостатистичні параметри можуть мати різну інтерпретацію.

Рис. 5. Приклад інформаційної довідки до індексу винятковості словника

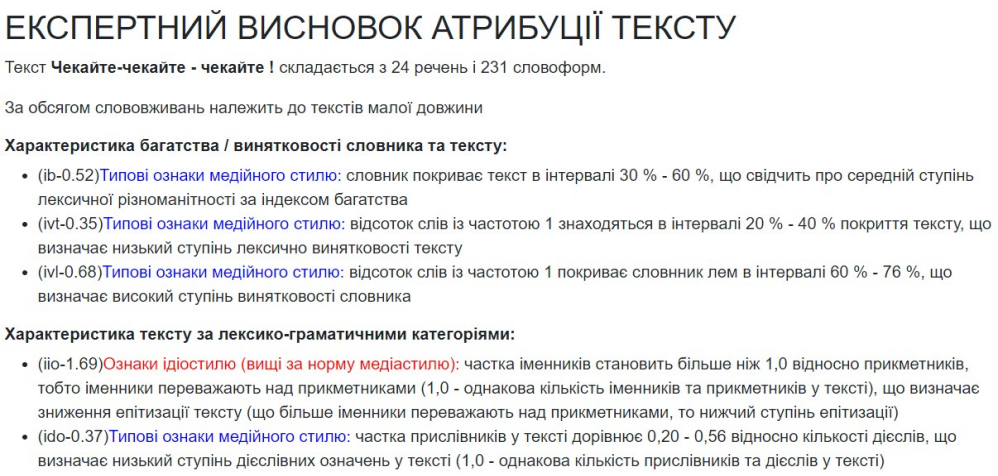
Рекомендація: зверніть увагу, що зіставлення числового значення статистичних параметрів уведеного тексту можливе лише для текстів медійного стилю, тому що на шкалі відкладено відрізок довірчого інтервалу лише медійного стилю; порівняльний аналіз текстів інших стилів позбавлений достовірності.

Як інтерпретувати експертний висновок атрибуції тексту? У рубриці "Експертний висновок атрибуції текстів" (рис. 6) подано експертний висновок про типові для медійного стилю чи індивідуальні лінгвостатистичні ознаки тексту. Цей висновок генерується автоматично за результатами зіставлення (рис. 4 – "Референс") обчислених числових значень індексів із пороговими значеннями довірчого інтервалу цього індексу в медійному стилі української мови.

Рекомендація: зверніть увагу, що експертний висновок достовірний лише для текстів медійного стилю.

Експертний висновок групує лінгвостатистичні висновки за кожним статистичним параметром за такими категоріями:

- характеристика багатства / винятковості словника та тексту;
- характеристика тексту за лексико-граматичними категоріями;
- характеристика розгортання подій тексту;
- характеристика психолінгвістичних ознак тексту;
- характеристика негативного сентименту тексту.


ЕКСПЕРТНИЙ ВИСНОВОК АТРИБУЦІЇ ТЕКСТУ

Текст **Чекайте-чекайте - чекайте !** складається з 24 речень і 231 словоформ.

За обсягом слововживань належить до текстів малої довжини

Характеристика багатства / винятковості словника та тексту:

- (ib-0.52)Типові ознаки медійного стилю: словник покриває текст в інтервалі 30 % - 60 %, що свідчить про середній ступінь лексичної різноманітності за індексом багатства
- (ivt-0.35)Типові ознаки медійного стилю: відсоток слів із частотою 1 знаходяться в інтервалі 20 % - 40 % покриття тексту, що визначає низький ступінь лексично винятковості тексту
- (ivl-0.68)Типові ознаки медійного стилю: відсоток слів із частотою 1 покриває словник лем в інтервалі 60 % - 76 %, що визначає високий ступінь винятковості словника

Характеристика тексту за лексико-граматичними категоріями:

- (iio-1.69)Ознаки ідіостилю (вищі за норму медіастилю): частка іменників становить більше ніж 1,0 відносно прикметників, тобто іменники переважають над прикметниками (1,0 - однакова кількість іменників та прикметників у тексті), що визначає зниження епітизації тексту (що більше іменники переважають над прикметниками, то нижчий ступінь епітизації)
- (ido-0.37)Типові ознаки медійного стилю: частка прислівників у тексті дорівнює 0,20 - 0,56 відносно кількості дієслів, що визначає низький ступінь дієслівних означень у тексті (1,0 - однакова кількість прислівників та дієслів у тексті)

Рис. 6. Фрагмент експертного висновку атрибуції тексту

Експертний висновок генерується по кожному індексу за варіантами відповіді на питання: Чи входить отримане числове значення індексу до довірчого інтервалу медійного стилю української мови? За результатами зіставлення числових значень із пороговими значеннями довірчого інтервалу можливі три відповіді:

1) Числове значення індексу входить до довірчого інтервалу медійного стилю української мови.

2) Числове значення індексу не входить до довірчого інтервалу медійного стилю української мови і нижче, ніж нижнє порогове значення цього інтервалу.

3) Числове значення індексу не входить до довірчого інтервалу медійного стилю української мови і вище, ніж верхнє порогове значення цього інтервалу.

Якщо за результатами автоматичного зіставлення (рис. 4 – "Референс") реалізовуються відповіді (2) і (3), то в такому випадку констатується лінгвостатистична

ознака ідіостилю автора. Три варіанти відповіді моделюють три експертні оцінки, які генеруються у системі трьома кольорами:

1. Типові ознаки медійного стилю – синій колір;
2. Ознаки ідіостилю (нижчі за норму медіастилю) – зелений колір;
3. Ознаки ідіостилю (вищі за норму медіастилю) – червоний колір.

Як інтерпретувати експертну оцінку? Кожен пункт експертного висновку подає експертну оцінку – лінгвостатистичну характеристику одного статистичного індексу. Числове значення статистичного параметра може отримати одну із трьох експертних оцінок, які формують прогностичний висновок про статистичну поведінку лінгвістичного явища за параметром у межах трьох діапазонів: 1) у межах довірчого інтервалу медіастилю; 2) у межах від нижнього порогового значення довірчого інтервалу медіастилю до 0; 3) у межах від вищого порогового значення довірчого інтервалу медіастилю до найвищого емпіричного значення індексу (у різних індексах це значення може бути різним). Для тих індексів, числове значення яких визначається відносно до кількості словоживань аналізованого тексту (обсягу тексту – 100 % словоживань), у лінгвостатистичному аналізі емпіричне значення інтерпретується через відсотковий еквівалент. Інформацію, систематизовану в експертній оцінці, можна диференціювати на чотири зони.

Приклад 1 експертної оцінки за індексом винятковості словника (індекс визначає, яку частку словника тексту становлять слова-гапакси, що зустрілися в тексті 1 раз). Згенерований системою текст: *(ivl-0.68) Типові ознаки медійного стилю: відсоток слів із частотою 1 покриває словник лем в інтервалі 60–76 %, що визначає високий ступінь винятковості словника.* Інтерпретація згенерованого тексту:

1) зона емпіричної інформації – (ivl-0.68): у дужках біля коду індексу винятковості словника (ivl) подано емпіричне числове значення цього індексу в аналізованому тексті – це відносна частота – 0,68 (емпіричні дані – ЕД).

2) зона експертної оцінки про індивідуальність / типовість стилю – (*Типові ознаки медійного стилю*). Використовуючи діагностичну модель для цього індексу – нижче / у межах / вище порогових показників довірчого інтервалу медіастилю (ЕД < 0,60–0,76 < ЕД), ми можемо зробити достовірний висновок про типові ознаки медійного стилю чи індивідуальні ознаки стилю автора. У прикладі значення індексу потрапляє в довірчий інтервал медійного стилю, тому в експертній оцінці подано характеристику *(Типові ознаки медійного стилю).*

3) зона статистичної інтерпретації емпіричного числового значення індексу (*відсоток слів із частотою 1 покриває словник лем в інтервалі 60–76 %*): інтервал інтерпретується у відсоткових значеннях (отримане числове значення множиться на 100–0,60 ×100=60 %) для кращого розуміння статистичної характеристики за індексом винятковості словника. Отримане числове значення індексу (0,68) є випадковим значенням (випадковою подією), тому що ми не можемо передбачити, якого числового значення набуде індекс у цьому чи в іншому тексті. Отже, у цій зоні подано прогностичну експертну оцінку про статистичну поведінку лінгвістичного явища в аналізованому тексті, але користувач може самостійно зіставити отримане значення індексу (0,68=68 %) із поданою характеристикою (пороговими значеннями довірчого інтервалу медійного стилю) і конкретизувати експертну оцінку: у наведеному прикладі частка слів із частотою 1 покриває 68 % словника.

4) зона експертної оцінки ступеня вияву статистичного параметра в аналізованому тексті та якісної інтерпретації лінгвістичних явищ (*що визначає високий ступінь винятковості словника*); у наведеному прикладі це оцінка ступеня винятковості / багатства словника для визначеного в попередній зоні інтервалу – це високий ступінь винятковості, а емпіричне значення індексу 0,68 конкретизує цей висновок: слова-гапакси покривають більше половини словника.

Для кращого розуміння того, як інтерпретувати експертні оцінки висновку атрибуції тексту, розглянемо приклади експертизи інших індексів.

Приклад 2 експертної оцінки за ступенем номінальності (індекс визначає відношення суми вживань іменників до суми вживань дієслів) тексту. Згенерований системою текст: *(stn-0.43) Ознаки ідіостилю (нижчі за норму медіастилю)* – (текст зеленим кольором)*: іменники становлять менш як 3,00 (високий ступінь номінальності тексту, іменники переважають відносно кількості дієслів у тексті), зниження індексу менше ніж 1,00 (1,00 – однакова кількість іменників і дієслів у тексті) засвідчує перевагу дієслів над іменниками й низький ступінь номінальності тексту.* Інтерпретація згенерованого тексту:

1) зона емпіричної інформації – (stn-0.43): у дужках біля коду індексу номінальності (stn) подано числове значення цього індексу в аналізованому тексті – 0,43 (емпіричні дані – ЕД).

2) зона експертної оцінки про індивідуальність / типовість стилю – *(Ознаки ідіостилю (нижчі за норму медіастилю)*. Використовуючи діагностичну модель для цього індексу – нижче / у межах / вище порогових показників довірчого інтервалу медіастилю (ЕД < 3–7,2 < ЕД), ми можемо зробити достовірний висновок про типові ознаки медійного стилю чи індивідуальні ознаки стилю автора. У прикладі значення індексу нижче, ніж нижнє порогове значення довірчого інтервалу медійного стилю (ЕД < 3), тому в експертній оцінці подано характеристику (*Ознаки ідіостилю (нижчі за норму медіастилю).*

3) зона статистичної інтерпретації емпіричного числового значення індексу (*іменники становлять менш як 3,00 (високий ступінь номінальності тексту, іменники переважають відносно кількості дієслів у тексті), зниження індексу менше ніж 1,00 (1,00 – однакова кількість іменників і дієслів у тексті) засвідчує перевагу дієслів над іменниками й низький ступінь номінальності тексту)*. Отримане числове значення індексу (0,43) є випадковим значенням (випадковою подією), тому що ми не можемо передбачити, якого числового значення набуде індекс у цьому чи в іншому тексті, і навіть у межах визначеного діапазону ЕД < 3. Отже, у цій зоні подано дві прогностичні експертні оцінки (дві підказки для користувача) про статистичну поведінку іменників щодо дієслів у діапазоні ЕД < 3: 1) за зниженням значення індексу до 1,0 *(1,0 – однакова кількість іменників і дієслів у тексті);* 2) за зниженням значення індексу нижче ніж 1,0. Користувач може самостійно зіставити отримане значення (0,43) із двома поданими статистичними характеристиками й конкретизувати експертну оцінку за питанням: "Наскільки переважає / не переважає частка іменників над дієсловами в аналізованому тексті?". У наведеному прикладі частка іменників не переважає над дієсловами і становить всього лише на 0,43 частки дієслів у тексті.

4) зона експертної оцінки ступеня вияву статистичного параметра в аналізованому тексті та якісної інтерпретації лінгвістичних явищ (*зниження відсотка до менше ніж 1,0 (1,0 – однакова кількість іменників та дієслів у тексті) засвідчує перевагу дієслів над іменниками)*;

у наведеному прикладі це оцінка ступеня номінальності тексту: маючи інформацію про лінгвістичну інтерпретацію зниження статистичного параметра за 1,0, користувач може зробити висновок про дуже низький ступінь номінальності тексту та значну перевагу в тексті дієслів, адже частка іменників становить лише 0,43 обсягу дієслів.

Приклад 3 експертної оцінки за індексом іменних означень тексту: індекс виражає відношення суми вживань іменників до суми вживань прикметників у тексті і свідчить про ступінь епітизації тексту: що менше іменників (що нижчий ступінь іменних означень), то вищий ступінь епітизації. Згенерований системою текст: *(ііо-1.69) Ознаки ідіостилю (вищі за норму медіастилю)* – (текст червоним кольором)*: частка іменників становить більше ніж 1,0 відносно прикметників, тобто іменники переважають над прикметниками (1,0 – однакова кількість іменників та прикметників у тексті), що визначає зниження епітизації тексту (що більше іменники переважають над прикметниками, то нижчий ступінь епітизації).* Інтерпретація згенерованого тексту:

1) зона емпіричної інформації – (ііо – 1,69): у дужках біля коду індексу іменних означень (ііо) подано числове значення цього індексу в аналізованому тексті – 1,69 (емпіричні дані - ЕД).

2) зона експертної оцінки про індивідуальність / типовість стилю – *(Ознаки ідіостилю (вищі за норму медіастилю)*. Використовуючи діагностичну модель для цього індексу – нижче / у межах / вище порогових показників довірчого інтервалу медіастилю (ЕД < 0,50–1,00 < ЕД), ми можемо зробити достовірний висновок про типові ознаки медійного стилю чи індивідуальні ознаки стилю автора. У прикладі значення індексу вище, ніж верхнє порогове значення довірчого інтервалу медійного стилю (1,00 < ЕД), тому в експертній оцінці подано характеристику *(Ознаки ідіостилю (вищі за норму медіастилю)*.

3) зона статистичної інтерпретації емпіричного числового значення індексу *(частка іменників становить більше ніж 100 % відносно прикметників, тобто іменники переважають над прикметниками (100 % – однакова кількість іменників і прикметників у тексті)*. Отримане числове значення індексу (1,69) є випадковим значенням (випадковою подією), тому що ми не можемо передбачити, якого числового значення набуде індекс у цьому чи в іншому тексті, і навіть у межах визначеного діапазону 1,00 < ЕД. Отже, у цій зоні подано прогностичну експертну оцінку про статистичну поведінку іменників щодо прикметників в аналізованому тексті, але користувач може самостійно зіставити отримане значення індексу (1, 69) із пороговим значенням довірчого інтервалу медійного стилю (1,00 < ЕД) і конкретизувати експертну оцінку за питанням: "Наскільки переважає / не переважає частка іменників над прикметниками в аналізованому тексті?". У наведеному тексті частка іменників переважає над прикметниками на 0,69 (1,69–1,00= 0,69), тому що при 1,00 кількість іменників і прикметників у тексті однакова.

4) зона експертної оцінки ступеня вияву статистичного параметра в аналізованому тексті та якісної інтерпретації лінгвістичних явищ *(тобто іменники переважають над прикметниками, що визначає зниження епітизації тексту (що більше іменники переважають над прикметниками, то нижчий ступінь епітизації)*; це лінгвістична інтерпретація ступеня вияву індексу іменних означень у тексті: у наведеному прикладі високий ступінь аналізованого індексу засвідчує низьку епітизацію тексту.

Рекомендація: зверніть увагу, що в інтерпретації отриманого числового значення індексу варто переводити емпіричне число індексу у відсотковий еквівалент лише у тих індексах, які у знаменнику мають числове значення обсягу слововживань тексту; ця заувага стосується таких індексів: індексу багатства тексту, індексу винятковості тексту, індекс винятковості словника, індексу модальності, індексу субстантивності, індексу прономіналізації, коефіцієнту агресивності

Як порівняти статистичні параметри двох текстів? У рубриці "Порівняння атрибуції текстів" є кнопка "Порахувати векторну відстань". Після активації цієї кнопки актуалізується інформація, систематизована в табличному форматі (рис. 7): 1 колонка – перелік усіх текстів, уведених користувачем, окрім того тексту, який аналізується системою в попередніх рубриках; 2 колонка – кількість речень у тексті; 3 колонка – кількість слів у тексті; 4 колонка – числове значення векторної відстані між аналізованим текстом та текстом рядка таблиці (список текстів у цій систематизації впорядковано за збільшенням числового значення векторної відстані); 5 колонка – опція "Порівняти", яка актуалізує автоматичне зіставлення всіх статистичних індексів аналізованого тексту з обраним у цій рубриці текстом через актуалізацію цієї опції у рядку обраного тексту.


ПОРІВНЯННЯ АТРИБУЦІЇ ТЕКСТІВ

Підрахувати векторну відстань

Назва тексту	Речень	Слів	Відстань	Порівняти
Гончаренко2-18.03.2024	20.00	141.00	0.44	Порівняти
Гончаренко4-17.03.2024	82.00	556.00	0.56	Порівняти
Гончаренко5-18.03.2024	25.00	210.00	0.77	Порівняти
ТРЕТЯ СВІТОВА ВІЙНА І ВИЗВОЛЬНА БОРОТЬБА	475.00	10,389.00	1.15	Порівняти
Люди з червоною ручкою	12.00	129.00	1.16	Порівняти
Короче мені сумно!	24.00	215.00	1.32	Порівняти
ДО ПРОБЛЕМИ ПОЛІТИЧНОЇ КОНСОЛІДАЦІЇ	355.00	8,561.00	1.54	Порівняти
З МОСКАЛЯМИ НЕМА СПІЛЬНОЇ МОВИ	58.00	1,406.00	1.78	Порівняти
Чому Путін одразу після "виборів" виступив на колегії ФСБ — аналіз ISW	21.00	439.00	2.37	Порівняти
Чекайте-чекайте - чекайте !	24.00	231.00	4.66	Порівняти

Рис. 7. Функція порівняння двох текстів за Евклідовою відстанню

Рекомендація: зверніть увагу, що для вірогідності статистичного аналізу в системі TextAttributor можна порівнювати між собою лише тексти приблизно однакового обсягу: короткі тексти з короткими текстами (обсягом до 1000 слововживань), а довгі тексти з довгими текстами (обсягом більше ніж 1000 слововживань).

Що таке векторна відстань? Векторна відстань між двома текстами обчислюється за формулою Евклідової відстані:

$$E(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{D} (A_i - B_i)^2},$$

де A_i і B_i – координати вектора, якими в системі TextAttributor виступають числові значення одного індексу у двох текстах, таким чином, обчислюється сума квадрата різниць між числовими значеннями 15 статистичних параметрів двох текстів, а потім із числа суми добувається корінь квадратний. Числові значення векторної відстані показують, наскільки тексти відрізняються статистично один від одного, яка міра відстані між цими текстами. Коливання значення векторної відстані для медійних текстів невідомі, але відстань може мати значення 0, якщо порівняти між собою один і той самий текст. Тобто тексти різні, якщо векторна відстань більше ніж 0. Що більше числове значення векторної відстані, то більше тексти відрізняються в лінгвостатистичному аспекті.

Як інтерпретувати індекси атрибуції в зіставленні двох текстів? Після актуалізації лінку "Порівняти" в рубриці "Індекси атрибуції тексту" (рис. 4, 8) оновлюється інформація та подаються систематизовані дані про зіставлення всіх статистичних індексів аналізованого тексту з текстом, обраним користувачем у рубриці "Порівняння атрибуції текстів".

ІНДЕКСИ АТРИБУЦІЇ ТЕКСТУ

	Гончаренко3-18.03.2024	Чекайте-чекайте - чекайте !	Референс
Кількість слів тексту	386	231	
Не оброблені системою слова	4	10	
Обсяг словника слів	202	115	
Кількість речень	39	24	
Статистичні параметри кількісного співвідношення слів реєстру словника та обсягу текстових слововживань:			
Індекс багатства	0.54	0.52	0.30 0.60
Індекс винятковості словника	0.66	0.68	0.60 0.76
Індекс винятковості тексту	0.36	0.35	0.20 0.40
Статистичні параметри кількісного співвідношення лексико-граматичних класів слів у тексті:			
Індекс іменних означень	1.07	1.69	0.50 1.00
Індекс дієслівних означень	0.57	0.37	0.20 0.56

Рис. 8. Фрагмент унаочненого порівняння статистичної параметризації двох текстів на шкалі довірчого інтервалу

Під час оновлення інформації в рубриці "Індекси атрибуції тексту" формується додаткова колонка числових значень статистичних індексів другого тексту: числові значення аналізованого тексту "*Гончаренко 3-18.03.2024*" подаються голубим кольором, а числові значення другого тексту "*Чекайте-чекайте-чекайте!*" – помаранчевим кольором. Також змінюється інформація і в колонці "Референс" (рис. 8): подано унаочнене зіставлення двох числових значень одного індексу на шкалі нижнього і верхнього порогових числових значень, які формують відрізок довірчого інтервалу медійного стилю. Місце числового значення індексу першого тексту позначено на шкалі синім перевернутим трикутником, а індексу другого текст – помаранчевим перевернутим трикутником. У межах довірчого інтервалу містяться типові числові значення для медійних текстів, а вихід за межі цього інтервалу свідчить про те, що в аналізованому тексті нетипове для медійного стилю числове значення індексів. Графічне представлення числових значень одного індексу двох текстів на шкалі різними кольорами дозволяє користувачеві швидко й ефективно порівняти статистичні характеристики одного лінгвістичного явища в різних текстових вибірках.

Як визначити ознаки токсичності тексту? Ступінь токсичності тексту визначено семантичним статистичним параметром – індексом токсичності тексту. Цей індекс (itox) обчислюється за частотою вживання в тексті негативної лексики, формула враховує різні класи слів з негативною тональністю: $Itox = (e + |K|\ (m + t)) / n \times 10$, де n – обсяг тексту; e – кількість слів-емоціогенів; m – кількість слів мови ворожнечі; t – кількість токсичних словосполучень; K – коефіцієнт, який дорівнює – 2, він посилює значення слів на означення мови ворожнечі і токсичних сполук, що на шкалі з п'яти розрядів (+2, +1, 0, –1, –2) відповідає – 2; множення на 10 введено до формули з метою збільшення числа для кращого унаочнення відрізка на шкалі зіставлення числового значення індексу з довірчим інтервалом медійного стилю. Інтерпретація індексу токсичності: якщо текст нейтральний і не містить зовсім ознак токсичності, тобто абсолютно позбавлений лексики з негативною семантикою, то індекс токсичності матиме значення 0. Встановлення довірчого інтервалу медійного стилю за індексом токсичності засвідчило, що типові числові значення цього індексу для коротких текстів знаходяться в інтервалі 0,1–0,7, а для довгих текстів – 0,2–0,9.

Лексико-семантична інтерпретація індексу токсичності подається в системі TextAttributor в окремій рубриці "Лінгвістична експертиза токсичності тексту" (рис. 9).

У цій рубриці подано кількісний аналіз і семантичну характеристику негативної лексики аналізованого тексту за семантичними категоріями: емоціогени з негативною семантикою – 5 слів; вульгаризми – 2 слова, сексизми – 2 слова. Також у цій рубриці можна проаналізувати вербалізацію отриманої характеристики: в аналізованому тексті слова з негативною семантикою виділені жирним шрифтом. Перепрошуємо за публікацію тексту токсичного змісту, але це зроблено з метою демонстрації можливостей системи.

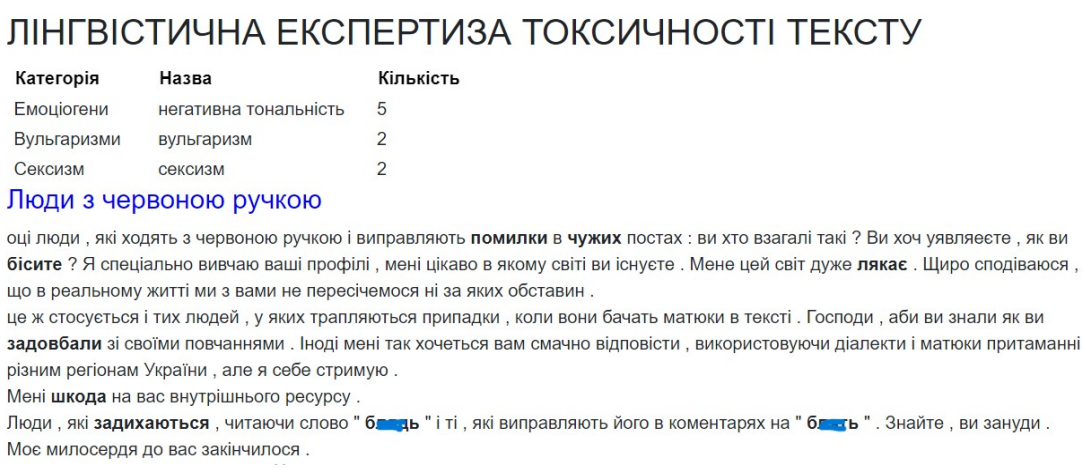


Рис. 9. Лінгвістична експертиза токсичності тексту "*Люди з червоною ручкою*"

Як інтерпретувати результати роботи нейронної мережі? До системи TextAttributor були імплементовані натреновані моделі нейронної мережі за двома завданнями: (а) визначення токсичності та (б) визначення авторства. У рубриці "Висновки нейронної мережі" подається результат роботи моделі глибокого машинного навчання (рис. 11). Ступінь токсичності тексту набуває значень від 0 (токсичності не виявлено) до 1 (найвища міра токсичності): текст *Чекайте-чекайте-чекайте!* має найвищий ступінь токсичності 1, за визначенням нейронної мережі. У завданні визначення авторства для відомих системі авторів, за текстами яких навчалася модель, обчислюється міра схожості тексту зі стилем певного автора, що набуває значень від 0 (схожість відсутня) до 1 (найбільша схожість). Відображаються результати аналізу тих тестів, які мають ступінь схожості більшу за 0.1. Текст *Чекайте-чекайте-чекайте!* має високий ступінь схожості з текстами О. Гончаренка – 0,91. Поточна версія системи працює в демонстраційному режимі та оперує невеликою кількістю текстів авторів, що виступають еталонними для визначення схожості введеного для аналізу тексту.

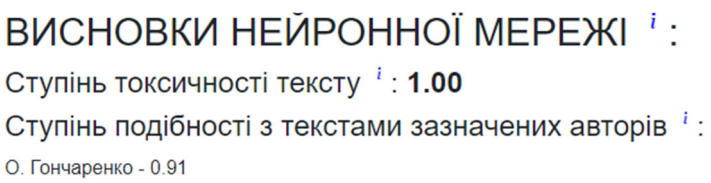


Рис. 10. Висновки нейронної мережі за текстом *Чекайте-чекайте-чекайте!*

Дискусія і висновки

Продемонстрований функціонал лінгвостатистичної системи TextAttributor 1.0 надає широкі аналітичні можливості не лише для наукового аналізу, а й для розв'язання широкого кола аналітичних завдань звичайного користувача інтернет-простору: верифікація авторства, психолінгвометричне профілювання, моделювання стилів, фільтрування текстової інформації в автоматичну моніторингу інтернет-простору, відстежування поширювачів токсичних текстів. Автоматична система перебуває на етапі тестування та вдосконалення описаних функцій, а також розбудови й поглиблення аналітичного функціоналу версії TextAttributor 1.1, зокрема натепер лінгвістична експертиза токсичності тексту доповнена визначенням фразеологізмів для тестування їхньої участі в маніпулятивних текстах, а також у системі додано розділ "Векторна візуалізація", у якому генерується наочне представлення атрибуції порівнюваних текстів (кількість текстів для порівняння задається користувачем) за чотирма статистичними параметрами методом UMAP (Uniform Manifold Approximation and Projection (McInnes, Healy, Melville 2020)), який дозволяє зменшити розмірність простору ознак шляхом проєкції на площину 15-вимірного простору (кожен статистичний параметр формує окремий вектор). До того ж, у розбудові системи TextAttributor колектив планує реалізувати інструмент статистичного порівняння текстів для всіх стилів української мови.

Колектив розробників системи TextAttributor бажає всім користувачам успішної роботи із застосуванням розроблених стилеметричних інструментів. Ми щиро сподіваємось, що наша система стане надійним помічником у виконанні ваших завдань. Будемо вдячні за ваші зауваження й рекомендації, які просимо надсилати за адресою автора публікації.

Подяки, джерела фінансування. Колектив розробників вебзастосунку TextAttributor 1.0. висловлює щиру подяку Уряду Великої Британії за фінансову підтримку проєкту "Визначення авторства анонімних українськомовних текстів з використанням методів штучного інтелекту в мережі Інтернет", у межах якого було створено цю систему, а також: директору громадської організації "Український науковий центр лінгвістичних студій" Костянтину Гончаренку за організацію цього проєкту, Світлані Яворській та Ірині Безкоровайній за підтримку проєкту і супровід у ході його реалізації, студентам четвертого курсу бакалаврської освітньої програми "Прикладна (комп'ютерна) лінгвістика та англійська мова" Київського національного університету імені Тараса Шевченка за допомогу у формуванні корпусу токсичних українськомовних текстів, випускниці нашої бакалаврської освітньої програми – Оксані ТОЛОЧКО за створення у межах бакалаврського проєкту словника тональної лексики української мови, який було використано для укладання лексикографічного списку негативних емоціогенів.

Список використаних джерел

Бук, С. (2021). *Велика проза Івана Франка: електронний корпус, частотні словники та інші міждисциплінарні контексти*. ЛНУ імені Івана Франка.

Визначення авторства анонімних українськомовних текстів у мережі Інтернет (2024). https://m.youtube.com/watch?v=JhGBn0l81uc

Зубань, О. (2020). *Комп'ютерне лексикографічне моделювання морфемної системи української мови*. ВПЦ "Київський університет". https://doi.org/10.17721/978-966-439-819-7

Перебийніс, В., Муравицька М., & Дарчук Н. (1985). *Частотні словники та їх використання*. Наукова думка.

Статистичні параметри стилів (1967). У В. С. Перебийніс (Ред.). *Наукова думка*.

TextAttributor: Інструкції користувачеві. (2024). http://ta.mova.info/instructions.

TextAttributor 1.0. (2024). http://ta.mova.info

Darchuk, N., Zuban, O., & Sorokin, V. (2024). On stylistic differentiation in ukrainian poetic speech based on the syntactic structure of sentences: *Bulletin of Taras Shevchenko National University of Kyiv. Literary Studies. Linguistics. Folklore Studies*, 1(35), 5–10. https://doi.org/10.17721/1728-2659.2024.35.01

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints 1802.03426. https://doi.org/10.48550/arXiv.1802.03426

References

Buk, S. N. (2021). *Long prose fiction by I. Franko: electronic corpus, frequency dictionaries and other interdisciplinary contexts.* Lviv Ivan Franko National University [in Ukrainian].

Determining the authorship of anonymous Ukrainian-language texts on the Internet (2024) [in Ukrainian]. https://m.youtube.com/watch?v=JhGBn0l81uc

Zuban, O. (2020). *Computer lexicographic modeling of the morpheme system of the Ukrainian language*. PPC "Kyiv University" [in Ukrainian]. https://doi.org/10.17721/978-966-439-819-7

Perebyinis, V., Muravytska, M., & Darchuk, N. (1985). *Chastotni slovnyky ta yih vykorystannia*. Naukova Dumka [in Ukrainian].

Statistical Parameters of Styles (1967). In V. S. Perebyinis (Eds.). Naukova Dumka [in Ukrainian].

TextAttributor 1.0. (2024) [in Ukrainian]. http://ta.mova.info

TextAttributor: User manual. (2024) [in Ukrainian]. http://ta.mova.info /instructions

Darchuk, N., Zuban, O., & Sorokin, V. (2024). On stylistic differentiation in ukrainian poetic speech based on the syntactic structure of sentences: *Bulletin of Taras Shevchenko National University of Kyiv. Literary Studies. Linguistics. Folklore Studies*, 1(35), 5–10. https://doi.org/10.17721/1728-2659.2024.35.01

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints 1802.03426. https://doi.org/10.48550/arXiv.1802.03426

Отримано редакцією журналу / Received: 17.06.24
Схвалено до друку / Accepted: 17.09.24

Oksana ZUBAN, PhD (Philol.), Assoc. Prof.
ORCID ID: 0000-0002-2644-3892
e-mail: oxana.zuban@knu.ua
Taras Shevchenko National University, Kyiv, Ukraine

FUNCTIONALITY OF THE AUTOMATIC SYSTEM FOR LINGUISTIC AND STATISTICAL ANALYTICS OF UKRAINIAN-LANGUAGE MEDIA TEXTS – TEXTATTRIBUTOR 1.0 (USER MANUAL)

The review article presents a user manual that introduces the functionality and operation of the automatic system for parameterizing Ukrainian-language media texts TextAttributor 1.0. This system is implemented as a web application (http://ta.mova.info/), which allows users to interactively perform a linguistic and statistical analysis of the input text and obtain statistical data on the parameterization of Ukrainian-language media text according to 18 statistical parameters. The system's functionality also generates an expert conclusion of the linguistic and statistical analysis of the text. It provides a graphical visualization of stylometric comparisons between one or two texts and the benchmark statistical characteristics of the Ukrainian-language media style. The individual modules of the system include 1) the Text Attribution Comparison module, which determines the degree of similarity between two texts selected by the user in the task of establishing authorship; 2) the Linguistic Expertise of Text Toxicity module, in which the user receives systematized linguistic and statistical data on the toxicity of a Ukrainian-language media text. The system is designed for researchers and general users interested in text information analytics to evaluate media texts in tasks such as authorship verification, psycholinguistic profiling, style modeling, text information filtering in automated internet monitoring, and tracking distributors of toxic texts. Free access to the TextAttributor 1.0 web application, a convenient interface, and the systematization of linguistic expert analytics of Ukrainian-language media texts provide users with broad opportunities to obtain the necessary information. The user manual was embedded on the website first, but now it is being published.

K e y w o r d s : ***Ukrainian-language media text, automatic linguistic analysis system, stylometry, text attribution, linguistic and statistical parameter, linguistic expertise, toxic text.***

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.