**Система TextAttributor як інструмент параметризації текстів**

Розвод Еліна Вадимівна, кандидат філологічних наук, доцент

Волинський національний університет імені Лесі Українки, Луцьк

ORCID ID: <https://orcid.org/0009-0000-3397-1688>

Сучасність змушує нас усвідомити важливість культури роботи з текстом, особливо в аспекті фільтрації, опрацювання та засвоєння закладеної в ньому інформації, яка іноді є результатом вторинної або третинної інтерпретації. Це також спонукає до осмислення тих механізмів, які не лише характеризують творче ставлення до мовленнєвого твору, а й можуть слугувати цілям навмисного викривлення фактологічних даних, історичної пам’яті та руйнування ціннісних орієнтирів особистості.

Параметризація – це процес виявлення референтного діапазону метрик обмеженої кількості параметрів для класифікації текстів певного типу, жанру або складності.

Американський лінгвіст Дуглас Байбер стверджує, що лінгвістичні параметри мають певний розподіл у текстах різних типів. Різнорівневі за складністю тексти та різними предметними областями можуть бути описані за допомогою достатнього набору параметрів, метрики яких є стабільними величинами [1].

У дослідженні розглянемо вебзастосунок TextAttributor 1.0, який розроблений фахівцями Київського національного університету імені Тараса Шевченка для автоматичного лінгвостатистичного аналізу українськомовних медійних текстів. Він дозволяє користувачам аналізувати тексти за різними параметрами: кількість слів, речень, обсяг словника та інші. Система також генерує експертний висновок та надає графічне унаочнення стилеметричного порівняння текстів з еталонними характеристиками українського медійного стилю.

До основних функцій  **TextAttributor 1.0 належать [3]:**

1. ***Лінгвостатистичний аналіз:*** система аналізує введений текст за 18 статистичними параметрами, надаючи детальні дані про його структуру та мовні особливості.
2. ***Експертний висновок:*** генерує експертний звіт на основі проведеного аналізу, що може бути корисним для дослідників та фахівців у галузі лінгвістики.
3. ***Стилометричне порівняння:*** дозволяє порівнювати один або два тексти з еталонними характеристиками медійного стилю української мови, що є корисним для встановлення авторства або аналізу стилю.
4. ***Модуль «Порівняння атрибуції текстів»:*** визначає ступінь схожості між двома обраними користувачем текстами, що допомагає у завданнях встановлення авторства.
5. ***Модуль «Лінгвістична експертиза токсичності тексту»:*** надає систематизовані лінгвістичні та статистичні дані про токсичність українськомовного медійного тексту.

Система обчислює 18 статистичних параметрів, з 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, до прикладу, діалектизми, русизми, оказіоналізми. Наступні 15 параметрів – статистичні індекси, згруповані таким чином [3]:

* 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;
	1. *за психолінгвістичними ознаками*;

 3) *за семантичною ознакою негативним сентиментом тексту*.

Актуальне дослідження ґрунтувалось на прикладі аналізу новинного дискурсу за такими параметрами: *кількісного співвідношення слів реєстру словника та обсягу текстових слововживань, кількісного співвідношення словосполучень та речень у тексті та* ***кількісного співвідношення лексико-граматичних класів слів у тексті.*** Результати аналізу тексту відображені у вигляді рисунків [2].

Рис.1.



Рис.2.

Статистичні параметри кількісного співвідношення слів реєстру словника

та обсягу текстових слововживань (Рис.2.):

1) індекс багатства (ib) – виражає відношення обсягу словника до обсягу тексту, свідчить про різноманітність словникового запасу; обернено пропорційний довжині тексту, тобто, що довший текст, то потенційно менше з’являється в ньому нових слів;

2) індекс винятковості тексту (ivt) – виражає відношення кількості слів-гапаксів, що зустрілися в тексті один раз, до обсягу тексту – свідчить про лексичне обмеження / лексичне багатство тексту;

3) індекс винятковості словника (ivl) – виражає відношення кількості слів-гапаксів, що зустрілися у тексті один раз, до обсягу словника – свідчить про лексичне обмеження / лексичне багатство словника.

Рис.3.

Статистичні параметри кількісного співвідношення словосполучень та речень у тексті (Рис.3.):

1) індекс динамічності (idyn) – виражає відношення суми дієслівних словосполучень (дієсл.+імен.) + (дієсл.+присл.) + (дієприсл.+імен.) + (дієприкм. + присл.) до суми іменних словосполучень (імен. + імен.)  + (прикм. + імен.), вказує на статичність/динамічність тексту, повільне/стрімке розгортання подій;

2) індекс зв’язності мовлення (izv) – виражає відношення суми кількості прийменників та сполучників до кількості речень у тексті,  свідчить про високий/нормальний/низький рівень зв’язності між описуваними подіями, явищами, ситуаціями.



**Рис. 4.**

**Статистичні параметри кількісного співвідношення лексико-граматичних класів слів у тексті (Рис.4.):**

**1) індекс іменних означень, або епітетизації, (iio) виражає відношення суми вживань іменників до суми вживань прикметників** – **свідчить про ступінь епітизації тексту: що менше іменників, то вищий ступінь епітизації;**

**2) індекс дієслівних означень (ido)** – **виражає відношення суми вживань прислівників до суми вживань дієслів** – **свідчить про ступінь вияву ознаки дії в тексті;**

**3) ступінь номінальності (stn)** – **виражає відношення суми вживань іменників до суми вживань дієслів** – **свідчить про ступінь номінальності тексту;**

**4) індекс прономіналізації (ipro)** – **виражає відношення кількості особових займенників до обсягу словоформ тексту, визначає ступінь кореферентності тексту;**

**5) індекс модальності (imod)** – **виражає відношення кількості часток до кількості слів у тексті** – **свідчить про ступінь емотивності тексту;**

**6) індекс субстантивності (isub)** – **виражає відношення кількості іменників до обсягу слововживань тексту** – **свідчить про насичення тесту іменниками, статичність стилю.**



Рис.5.

Отримані результати чітко демонструють (Рис.5.), що моніторинг концептуального простору є важливим для виявлення нових форм деструкції в масмедіа, а також для створення вдосконалених систем для роботи з потенційно небезпечними текстами із застосуванням нейромереж і машинного навчання. Запропонована параметризація медіадискурсу дозволяє виявити та здійснити комплексну ймовірну оцінку можливих ризиків щодо потенційних цілей інформаційних атак.

Література

1. Susan Conrad, Douglas Biber. Register, Genre, and Style. Cambridge: Cambridge University Press, 2019. 420p.
2. TextAttributor 1.0. (2024) [in Ukrainian]. <http://ta.mova.info> (date of application 10.01.2025).
3. TextAttributor: User manual. (2024) [in Ukrainian]. <http://ta.mova.info/instructions>