



І. Ю. Хомицька, В. М. Теслюк, І. Б. Базилевич

Національний університет "Львівська політехніка", м. Львів, Україна

ЕФЕКТИВНІСТЬ СТАТИСТИЧНИХ КРИТЕРІЇВ ДЛЯ ВИЗНАЧЕННЯ СТИЛЬОВИХ ХАРАКТЕРИСТИК ТЕКСТІВ

Проаналізовано особливості підвищення достовірності встановлення звичайної та стильової атрибуції текстів, застосовуючи статистичні критерії Лемана-Розенблатта та Стюдента. Запропоновано методологію застосування поєднання статистичних критеріїв Лемана-Розенблатта та Стюдента для здійснення атрибуції текстів на підставі визначення довжини слів та стильової атрибуції текстів на фонологічному рівні. Визначено залежність ефективності статистичних критеріїв від трьох основних чинників: обсягу вибірки, рівня мови, на якому здійснюється атрибуція текстів, та лінгвістичного матеріалу дослідження. Наведено алгоритм здійснення атрибуції текстів та стильової атрибуції текстів мовою програмування Java. Запропоновано поєднання двох основних підходів до здійснення стильової атрибуції текстів, яке може використовуватися і для авторської атрибуції. Перший підхід ґрунтується на визначенні розподілу довжини слів за допомогою критерію Лемана-Розенблатта, другий – на визначенні відмінностей у частоті вживання груп приголосних фонем, використовуючи критерій Стюдента. Встановлено оптимальний обсяг вибірки. У першому випадку обсяг вибірки становить 50 000 літер, у другому – 30 000 фонем. Проведено атрибуцію текстів на матеріалі статистичних даних, отриманих з художніх творів Данієля Дефо "Робінзон Крузо" (XVIII ст.) та Пауліни Сари Джо Моес "Я перед Вами" (XXI ст.). Стильову атрибуцію проведено на матеріалі статистичних даних, отриманих з газетного (The Daily Telegraph) (XXI ст.) та наукового стилів (Acta Physica Polonica) (XXI ст.). Встановлено ефективність статистичних критеріїв Лемана-Розенблатта та Стюдента для здійснення атрибуції та стильової атрибуції текстів, визначаючи розподіл довжини слів та частоту вживання груп приголосних фонем. Результати отримано з довірчою ймовірністю 95 %. Розроблено структуру програмної системи атрибуції тексту та стильової атрибуції. Структура ґрунтується на модульному принципі. Введено модулі атрибуції текстів та стильової атрибуції. Розроблене програмне забезпечення має широкий спектр застосування. Його можна використовувати і для авторської атрибуції. Практичне застосування атрибуції текстів та стильової атрибуції здійснюється у галузі філології та у галузях, що стосуються визначення атрибуції текстів та стильової приналежності тексту. Авторська атрибуція має практичну цінність у кожній галузі людської діяльності, в якій потрібно встановити авторство тексту. Це дослідження є розширенням і доповненням щодо наших попередніх досліджень, присвяченим авторській атрибуції.

Ключові слова: атрибуція текстів; стильова атрибуція текстів; розподіл довжини слів; частота вживання груп приголосних фонем; критерій Лемана-Розенблатта; критерій Стюдента.

Вступ / Introduction

До наукових напрямів, які широко дискутують й активно досліджують дослідники, входить застосування методів обчислювального інтелекту, а саме: статистичних методів для докладнішої характеристики взаємозв'язку лінгвістичних і статистичних характеристик тексту (атрибуція текстів); стильових характеристик текстів (стильова атрибуція текстів); авторських характеристик текстів (авторська атрибуція текстів) [3, 7]. Це дослідження є продовженням попередніх досліджень авторів статті, присвячених авторській атрибуції текстів [5]. Потреба опрацювання великих обсягів інформації передбачає виконання зазначених завдань. Актуальність дослідження полягає у визначенні ефективних

статистичних методів для забезпечення атрибуції текстів з високим рівнем точності. Тому потрібно визначити залежність ефективності статистичного критерію від обсягу вибірки, рівня мови (фонологічний, лексичний, синтаксичний) та лінгвістичного матеріалу.

Об'єкт дослідження – процес визначення статистичних характеристик англomовних текстів різних століть.

Предмет дослідження – методи та засоби автоматизації встановлення статистичними методами подібності та відмінності між цими характеристиками.

Мета роботи – підвищити достовірність встановлення звичайної та стильової атрибуції текстів, застосовуючи статистичні критерії Лемана-Розенблатта та Стюдента.

Інформація про авторів:

Хомицька Ірина Юріївна, канд. техн. наук, доцент, кафедра прикладної лінгвістики. Email: iryna.khomytska@ukr.net;

<https://orcid.org/0000-0003-3470-7197>

Теслюк Василь Миколайович, д-р техн. наук, професор, кафедра автоматизованих систем управління.

Email: vasyt.m.leslyuk@lpnu.ua; <https://orcid.org/0000-0002-5974-9310>

Базилевич Ірина Богданівна, канд. фіз.-мат. наук, доцент, кафедра математичної статистики та диференціальних рівнянь.

Email: i_bazylevych@yahoo.com

Цитування за ДСТУ: Хомицька І. Ю., Теслюк В. М., Базилевич І. Б. Ефективність статистичних критеріїв для визначення стильових характеристик текстів. Науковий вісник НЛТУ України. 2023, т. 33, № 4. С. 90–94.

Citation APA: Khomytska, I. Yu., Teslyuk, V. M., & Bazylevych, I. B. (2023). Efficiency of statistical tests for determining style characteristics of texts. *Scientific Bulletin of UNFU*, 33(4), 90–94. <https://doi.org/10.36930/40330413>

Для досягнення зазначеної мети визначено такі основні завдання дослідження: визначити обсяг вибірки, що дає змогу отримати достовірні результати для критеріїв Лемана-Розенблатта та Стюдента; встановити ефективність критеріїв для розподілу довжини слів та частоти вживання фонем; підібрати оптимальний лінгвістичний матеріал для забезпечення достовірних результатів.

Аналіз останніх досліджень та публікацій. Актуальними проблемами сьогодення є автоматизоване виявлення підроблення документів (показів свідків, заповітів); неправдивих відомостей у повідомленнях; образливих, погрозових і компрометуючих матеріалів; встановлення авторства наукових статей та інших публікацій, які мають більш ніж одного автора. Вивченню зазначених проблем присвячена низка досліджень, основними напрямками яких є застосування методів машинного навчання та класичних методів для встановлення авторства текстів, або здійснення авторської атрибуції [3, 7]. Статистичні параметри англійської поезії розглядають як критерії визначення авторського стилю. Авторський стиль визначається на лексичному рівні [8, 9]. Застосування методів математичної статистики для розв'язання поставленої задачі є ефективним. Проаналізовано сучасні підходи до встановлення авторства, які різняться вибором статистичного критерію, рівня мови (фонологічного, лексичного, синтаксичного) та матеріалом дослідження. Відповідно, актуальним завданням сьогодення є підвищення точності атрибуції текстів та стильової атрибуції текстів.

Наведемо результати зіставлення логістичної регресії та класифікаторів Байеса для визначення негативного змісту повідомлень у соціальних мережах. З'ясовано, що кращі результати дає логістична регресія – 91,1 % [1]. Матриці неподібності використовували для ідентифікації авторського стилю на лексичному рівні. Максимальна отримана точність – 98,75 % [2]. Метод нейронних мереж забезпечив точність вищу ніж 90 % на матеріалі газетного стилю у соціальних мережах, внаслідок чого визначено позитивну та негативну тональність авторського стилю [11]. У зазначених дослідженнях задачу підвищення рівня точності розв'язано частково. Для розв'язання цієї задачі потрібно встановити метод чи поєднання методів, які забезпечать високий рівень точності на оптимально вибраному рівні мови. У нашому дослідженні пропонуємо застосувати поєднання статистичних критеріїв Лемана-Розенблатта та Стюдента для визначення розподілу довжини слів та на фонологічному рівні. Рівень значущості вибрано 5 %. Результати отримано з довірчою ймовірністю 95 %.

Матеріали та методи дослідження. У дослідженні застосовано такі методи: статистичний – метод гіпотез; метод статистичного моделювання – для розмежування статистичних структур стилів; метод теорії алгоритмів та метод об'єктно-орієнтованого підходу – для розроблення програмного засобу [5]. Атрибуцію текстів проведено на матеріалі статистичних даних, отриманих у спосіб розмежування статистичних структур авторських стилів Данієля Дефо ("Робінзон Крузо", XVIII ст.) та Пауліни Сари Джо Моєс ("Я перед Вами" XXI ст.). Обсяг вибірки становить 50 000 літер. Застосований метод гіпотез містить використання статистичного критерію Лемана-Розенблатта та статистичного критерію Стюдента. Статистичний критерій Лемана-Розенблатта

та застосовано для розмежування текстів за визначенням розподілу довжини слів. Стильову атрибуцію проведено на матеріалі статистичних даних, отриманих шляхом диференціації статистичних структур з газетного стилю (The Daily Telegraph, XXI ст.) та наукового стилю (Acta Physica Polonica, XXI ст.). Обсяг вибірки становить 30 000 фонем. Статистичний критерій Стюдента використано на фонологічному рівні. Зазначені критерії визначено для перевірки гіпотези на однорідність – чи зіставлявані вибірки є вибірками з одного розподілу. У дослідженні вибрано рівень значущості 0,05. Результати отримано з довірчою ймовірністю 0,95.

Результати дослідження та їх обговорення / Research results and their discussion

Результати дослідження отримано, відповідно до алгоритму застосування статистичних критеріїв для атрибуції текстів та стильової атрибуції текстів:

Крок 1. Об'єднати елементи вибірки з творів Д. Дефо (перша вибірка) та П. Моєс (друга вибірка).

Крок 2. Об'єднати елементи вибірки з наукового стилю (перша вибірка) та газетного стилю (друга вибірка).

Крок 3. Упорядкувати вибірки за зростанням.

Крок 4. Визначити ранги для кожної з упорядкованих вибірок.

Крок 5. Застосувати статистику критерію Лемана-Розенблатта [4, 6, 10]:

$$\omega_{n,m}^2 = \int_{-\infty}^{\infty} [\hat{F}_n(x) - G_m(x)]^2 d\hat{H}_{n+m}(x),$$

де: $\hat{F}_n(x)$ і $G_m(x)$ – емпіричні функції розподілу вибірок

обсягами n і m ; $\hat{H}_{n+m}(x) = \frac{n}{n+m} \hat{F}_n(x) + \frac{m}{n+m} \hat{G}_m(x)$ – емпірична функція об'єднаної вибірки. Значення статистики залежить тільки від рангів елементів вибірки:

$$\omega_{n,m}^2 = \frac{1}{nm} \left[\frac{1}{6} + \frac{1}{m} \sum_{i=1}^n (R_i - i)^2 + \frac{1}{n} \sum_{j=1}^m (S_j - j)^2 \right] - \frac{2}{3},$$

де: R_i – ранг $x_{(i)}$; S_j – ранг $y_{(j)}$ в об'єднаній варіаційній вибірці.

Крок 6. Перевірити чи відмінності у вибіркових показниках є статистично значущими.

Крок 7. Обчислити значення критерію χ^2 за формулою [4, 6, 10]: $\chi^2 = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i'}$, де: k – кількість інтервалів; n_i – кількість частот вживання фонем, які потрапляють в i -й інтервал, n_i' – теоретичне значення частоти.

Крок 8. Застосувати критерій Стюдента [4, 10]: $t = \frac{\bar{x} - \bar{x}_0}{S} \sqrt{n}$, де: $\bar{x}$ – середнє значення частоти груп фонем, отримане з вибірки, з якою ми працюємо; n – кількість порцій цієї вибірки; $\bar{x}_0$ – середнє значення частоти групи приголосних фонем генеральної сукупності; S^2 – незміщена оцінка дисперсії. Статистика t підлягає розподілу Стюдента [10].

Крок 9. Перевірити чи відмінності у вибіркових показниках є статистично значущими.

Розроблено програмну систему атрибуції тексту та стильової атрибуції тексту мовою програмування Java [5]. Структура системи ґрунтується на модульному принципі (модуль введення/виведення даних; модуль формування вибірки; модуль обчислення мовних одиниць; модуль застосування статистичних критеріїв). Додано модулі атрибуції текстів та стильової атрибуції.

Алгоритм функціонування системи атрибуції текстів та стильової атрибуції є такий:

Крок 1. Відбір текстів з художнього, газетного та наукового стилів.

Крок 2. Створення транскрипційних варіантів для художнього стилю.

Крок 3. Формування вибірки з приголосних фонем обсягом 30 000 фонем.

Крок 4. Визначення середнього значення частоти вживання приголосних фонем.

Крок 5. Застосування критерію Стюдента для здійснення атрибуції текстів.

Крок 6. Створення вибірки з лексичних одиниць з газетного та наукового стилів обсягом 50 000 літер.

Крок 7. Визначення довжини слів у текстах з газетного та наукового стилів.

Крок 8. Застосування критерію Лемана-Розенблатта для стильової атрибуції текстів.

Статистичний критерій Лемана-Розенблатта використано, визначаючи розподіл довжини слів на матеріалі статистичних даних, отриманих з художніх творів Данієля Дефо "Робінзон Крузо" (XVIII ст.) та Пауліни Сари Джо Моєс "Я перед Вами" (XXI ст.). Обсяг вибірки становить 50 000 літер. Для здійснення атрибуції текстів перевірено нульову гіпотезу однорідності H_0 . У табл. 1 наведено емпіричні дані та ранги, у табл. 2 – упорядковані ранги та квадрати різниць.

Табл. 1. Емпіричні дані та ранги для атрибуції текстів Д. Дефо і П. Моєс / Empirical data and ranks for attribution of texts by D. Defoe and P. Moyes

i, j	X	Y	R_i	S_j
1	6	4	33	17,5
2	7	4	36,5	17,5
3	4	5	17,5	27
4	5	4	27	17,5
5	4	1	17,5	1
6	5	5	27	27
7	3	5	8	27
8	6	3	33	8
9	7	3	36,5	8
10	3	6	8	33
11	7	2	36,5	2,5
12	3	3	8	8
13	5	4	27	17,5
14	4	3	17,5	8
15	4	7	17,5	36,5
16	3	5	8	27
17	5	3	27	8
18	4	2	17,5	2,5
19		4		17,5
20		5		27

Розрахунки: суми квадратів різниць становить відповідно $\sum_{i=1}^n (R_i - i)^2 = 3481,00$ і $\sum_{j=1}^n (S_j - j)^2 = 1335,00$; обсяги вибірок: $n = 18$ і $m = 20$; критерій $\omega^2 = 0,02$; параметри $M = 0,17$; $D = 0,02$; значення емпіричної нормованої статистики $Z \approx 0,22$. Квантиль $Z_{0,95} = 0,46$.

Оскільки $Z < Z_{0,95}$ ($0,22 < 0,46$), нульова гіпотеза H_0 приймається на рівні значущості 0,05. Відмінності між зіставляваними вибірками не є статистично значущі. Отже, здійснено атрибуцію текстів, зокрема встановлено, що тексти художніх творів Данієля Дефо "Робінзон Крузо" (XVIII ст.) та Пауліни Сари Джо Моєс "Я перед Вами" (XXI ст.) написані англійською мовою.

Статистичний критерій Стюдента використано для стильової атрибуції, перевіряючи гіпотезу на однорід-

ність H_0 – чи зіставлявані вибірки є вибірками з одного розподілу. Стильову атрибуцію здійснено на фонологічному рівні, на матеріалі статистичних даних, отриманих з газетного (The Daily Telegraph) (XXI ст.) та наукового стилів (Acta Physica Polonica) (XXI ст.). Обсяг вибірки становить 30 000 фонем.

Табл. 2. Упорядковані ранги та квадрати різниць для атрибуції текстів Д. Дефо і П. Моєс / Ordered ranks and squared differences for attribution of texts by D. Defoe and P. Moyes

i, j	R_i	S_j	$(R_i - i)^2$	$(S_j - j)^2$
1	8	1	49,00	0,00
2	8	2,5	36,00	0,25
3	8	2,5	25,00	0,25
4	8	8	16,00	16,00
5	17,5	8	156,25	9,00
6	17,5	8	132,25	4,00
7	17,5	8	110,25	1,00
8	17,5	8	90,25	0,00
9	17,5	17,5	72,25	72,25
10	27	17,5	289,00	56,25
11	27	17,5	256,00	42,25
12	27	17,5	225,00	30,25
13	27	17,5	196,00	20,25
14	33	27	361,00	169,00
15	33	27	324,00	144,00
16	36,5	27	420,25	121,00
17	36,5	27	380,25	100,00
18	36,5	27	342,25	81,00
19		33		196,00
20		36,5		272,25

У табл. 3 наведено дані для восьми груп фонем (Гр.ф.); середні частоти і суми квадратів різниць для наукового стилю (HC) та газетного стилю ($ГС$).

Табл. 3. Середні частоти та суми квадратів різниць для стильової атрибуції HC та $ГС$ / Mean frequencies and sums of squares of differences for style attribution of SS and NS

Гр.ф.	$HC \bar{x}$	$\Sigma(x_i - \bar{x})^2$	$ГС \bar{x}$	$\Sigma(x_i - \bar{x})^2$
1	123,5	5465,75	130,1	6565,51
2	425,1	11442,71	406,4	4991,76
3	8,4	419,36	7,3	1929,59
4	60,9	2282,71	64,7	2702,39
5	87,9	3125,91	82,3	1450,79
6	220,4	6461,16	225,9	4517,51
7	186,8	2872,24	183,4	1851,36
8	211,1	4202,71	204,5	4025,75

У табл. 4 наведено значення дисперсії S значення величини t ; при оцінці різниць типу $\bar{x}_1 - \bar{x}_2$ застосовується двосторонній рівень значущості $2Q$. Тип значення величини $\bar{x}_1 - \bar{x}_2$ є або випадковий, або істотний. Істотні відмінності встановлено для груп передньоязикових, носових і зімкнених фонем.

Табл. 4. Визначення істотних відмінностей між науковим і газетним стилями за критерієм Стюдента / Determining essential differences between the scientific and newspaper styles by the Student's t-test

Гр.ф.	S	t	$2Q, \%$	$\bar{x}_1 - \bar{x}_2$
1	14,16	1,84	5	випадковий
2	16,55	4,44	<0,1	істотний
3	6,26	0,69	>20	випадковий
4	9,12	1,64	>10	випадковий
5	8,73	2,52	<2	істотний
6	13,52	1,60	>10	випадковий
7	8,87	1,51	>10	випадковий
8	11,71	2,22	<5	істотний

Проведене дослідження дає змогу ствердити, що статистичний критерій Лемана-Розенблатта є ефективний для атрибуції текстів, визначаючи розподіл довжини слів. Критерій Стюдента є ефективний для стильової атрибуції текстів на фонологічному рівні.

Обговорення результатів дослідження. Запропоноване поєднання статистичного критерію Лемана-Розенблатта та критерію Стюдента для встановлення авторства тексту, дає кращі результати (95 %), ніж методи логістичної регресії (91,1 %) та класифікаторів Байєса [1]. Для встановлення авторства у наукових статтях, які мають більше ніж одного автора та у повідомленнях негативного змісту (погрозового, образливого) в останніх дослідженнях використовуються методи машинного навчання [3]. У дослідженнях цього напрямку, метод штучних нейронних мереж дає максимальну точність – 98,75 % [2]. Однак, у більшості випадків, при встановленні авторства за послідовністю слів, літер, точність є вищою ніж 90 % [7, 11]. Зазначений рівень точності встановлено для авторської атрибуції за частотою вживання слів у поетичних творах чеської, німецької, іспанської та англійської мов, а також за частотою одиниць синтаксичного рівня в англомовних текстах [8, 9].

Наприклад, у роботі [9] проаналізовано різні лінгвістичні аспекти, які можуть допомогти представити стиль тексту. Зокрема, автори вивчають роль синтаксису та лексичних слів (іменників, дієслів, прикметників та прислівників) у поданні стилю. Вони використовують суто синтаксичну мовну модель для вивчення значення структур речень як в однодомній, так і в міждомній атрибуції, тобто міжтематичною та міжжанровою атрибуцією. Автори показали, що синтаксис може бути корисним для міжжанрової атрибуції, тоді як для міжтематичної атрибуції та однодомної може бути корисна додаткова лексична інформація. Окрім цього, чисті синтаксичні моделі можуть бути неефективними самі по собі і їх потрібно використовувати в поєднанні з іншими надійними моделями.

Поєднання критерію Лемана-Розенблатта та критерію Стюдента дає змогу отримати результати з довірчою ймовірністю 95 % для розподілу довжини слів і на фонологічному рівні для атрибуції текстів та стильової атрибуції. Зазначене поєднання критеріїв можна використати і для авторської атрибуції.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – запропоновано методологію на підставі поєднання статистичних критеріїв Лемана-Розенблатта та Стюдента для здійснення атрибуції текстів та стильової атрибуції текстів.

Практична значущість результатів дослідження – отримані результати можна застосувати для встановлення авторства документів, стильової характеристики художньої літератури, визначення авторського стилю наукових статей, і загалом, у кожній галузі людської діяльності, в якій потрібно встановити стильові та авторські характеристики тексту.

Висновки / Conclusions

Розроблено методологію у спосіб поєднання статистичних критеріїв Лемана-Розенблатта та Стюдента для здійснення атрибуції текстів на підставі визначення

довжини слів та стильової атрибуції текстів на фонологічному рівні. Визначено залежність ефективності статистичних критеріїв від трьох основних чинників: обсягу вибірки, рівня мови, на якому здійснюється атрибуція текстів та лінгвістичного матеріалу дослідження.

Подано алгоритм реалізації атрибуції текстів та стильової атрибуції текстів мовою програмування Java. У наявну структуру програмної системи введено модулі атрибуції текстів та стильової атрибуції. Розроблене програмне забезпечення в подальших дослідженнях можна використовувати і для авторської атрибуції.

Атрибуцію текстів здійснено на матеріалі статистичних даних, отриманих з художніх творів Данієля Дефо "Робінзон Крузо" (XVIII ст.) та Пауліни Сари Джо Моєс "Я перед Вами" (XXI ст.). Стильову атрибуцію текстів проведено на матеріалі статистичних даних, отриманих з газетного (The Daily Telegraph) (XXI ст.) та наукового стилів (Acta Physica Polonica) (XXI ст.). У дослідженні досягнуто поставленої мети – підвищено достовірність встановлення звичайної та стильової атрибуції текстів, застосовуючи статистичні критерії Лемана-Розенблатта та Стюдента для визначення розподілу довжини слів та частоти вживання груп приголосних фонем. Результати отримано з довірчою ймовірністю 95 %.

References

1. Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and Naive Bayes classifiers, in: 2018 international conference on information reuse and integration, 269–276. <https://doi.org/10.1109/IRI.2018.00049>
2. Akimushkin, C., Amancio, D. R., & Oliveira, O. N. (2018). On the role of words in the network structure of texts: Application to authorship attribution, Physica A: Statistical Mechanics and its Applications, 495, 49–58. <https://doi.org/10.1016/j.physa.2017.12.054>
3. Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatos, E., Stein, B., Wiegmann, M., & Zangerle, E. (2020). Shared Tasks on Authorship Analysis at PAN 2020. In book: Advances in Information Retrieval, 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, Proceedings, Part II, 508–516. http://doi.org/10.1007/978-3-030-45442-5_66
4. Bhattacharya, R., & Waymire, E. C. (2017). A Basic Course in Probability Theory. Textbook. Springer, 2nd ed. <https://doi.org/10.1007/978-0-387-71939-9>
5. Khomytska, I., Teslyuk, V., Kryvinska, N., & Bazylevych, I. (2020). Software-Based Approach Towards Automated Authorship Acknowledgement – Chi-Square Test on One Consonant Group. July Electronics, 9(7), 1138. <http://doi.org/10.3390/electronics9071138>
6. Kornai, A. (2008). Mathematical Linguistics. Springer. Textbook. <https://doi.org/10.1007/978-1-84628-986-6>
7. Muttenthaler, L., Lucas, G., & Amann, J. (2019). Authorship Attribution in Fan-Fictional Texts Given Variable Length Character and Word N-Grams. Notebook for PAN at CLEF 2019, 9–12 September, Lugano, Switzerland, 2380, 49. URL: https://ceur-ws.org/Vol-2380/paper_49.pdf
8. Plechac, P., Bobenhausen, K., & Hammerich, B. (2018). Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry, Studia Metrica et Poetica, 5(2), 29–54. <http://doi.org/10.12697/smp.2018.5.2.02>
9. Sundararajan, K., & Woodard, D. (2018). What represents "style" in authorship attribution?. In Proceedings of the 27th International Conference on Computational Linguistics, 2814–2822, Santa Fe, New Mexico, USA. Association for Computational Linguistics. URL: <https://aclanthology.org/C18-1238>

10. Turchyn, V. (1999). Matematychna statystyka. Navch. Posib. Vydavnychj tsentr "Akademia": Kyiv, Ukraine. [In Ukrainian].
11. Voloshyn, S., Vysotska, V., Markiv, O., Dyyak, I., Budz, I., & Schuchmann, V. (2022). Sentiment Analysis Technology of English Newspapers Quotes Based on Neural Network as Public

Opinion Influences Identification Tool. 2022 17th International Conference on Computer Sciences and Information Technologies (CSIT), 83–88, <http://doi.org/10.1109/CSIT56902.2022.10000627>

I. Yu. Khomytska, V. M. Teslyuk, I. B. Bazylevych

Lviv Polytechnic National University, Lviv, Ukraine

EFFICIENCY OF STATISTICAL TESTS FOR DETERMINING STYLE CHARACTERISTICS OF TEXTS

In the research, we have analyzed the peculiarities of enhancing the test validity of authorship attribution and style attribution using the Lehmann-Rosenblatt and the Student's t statistical tests. The novelty of the research is a proposed approach of the use of the Lehmann-Rosenblatt and the Student's t-tests to attribution of texts determining the word length and style attribution of texts on the phonological level. It is topical to determine the dependence of the statistical tests efficiency on following three main factors: the sample size, the language level, on which the attribution is done, and the linguistic material under study. We have presented the algorithm of text processing on the Java programming language. We propose a combination of two main approaches to attribution of texts and style attribution of texts. It is possible to use these approaches for authorship attribution as well. The first approach is based on determining word length distribution by the Lehmann-Rosenblatt test. In the second approach, we use the Student's t-test for determining the differences in frequencies of occurrence of consonant phoneme groups. We have established an optimal sample size. It is 50 000 letters in the first case and 30 000 phonemes in the second case. The attribution of texts has been done on the material of statistical data for the literary works "Robinson Crusoe" by Daniel Defoe (18th century) and "Me before You" by Pauline Sara Jo Moyes (21st century). Style attribution has been performed on the material of statistical data for the newspaper style (The Daily Telegraph) (21st century) and the scientific style (Acta Physica Polonica) (21st century). The goal of the research has been met— the Lehmann-Rosenblatt and the Student's t-tests have proved to be efficient for attribution of texts and style attribution based on word length distribution and frequencies of occurrence of consonant phoneme groups. The test validity of the obtained results is 95 %. The structure of the software for attribution of texts and style attribution has been developed. The structure is based on a modular principle. Two more modules, in particular, modules for attribution of texts and style attribution have been introduced to the structure. The developed software has a wide range of applications. It may be used for authorship attribution as well. Attribution of texts and style attribution can be practically applied in philology and other fields related to the problem. Practical value of authorship attribution is in every sphere of human activity where it is necessary to identify the author. If compared with our previous research on authorship attribution, this one is expanded and supplemented.

Keywords: attribution of texts; style attribution of texts; word length distribution; frequencies of occurrence of consonant phoneme groups; the Lehmann-Rosenblatt test; the Student's t-test.