Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.36930/40300522
Article received 03.11.2020 p.
Article accepted 23.10.2020 p.
UDC 004.9

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

@ Correspondence author
V. M. Teslyuk
vasyl.teslyuk@gmail.com

І. Ю. Хомицька[1], В. М. Теслюк[1], І. Б. Базилевич[2], В. В. Береговський[3]

[1] Національний університет "Львівська політехніка", м. Львів, Україна
[2] Львівський національний університет ім. Івана Франка, м. Львів, Україна
[3] Івано-Франківський національний технічний університет нафти і газу, м. Івано-Франківськ, Україна

СТАТИСТИЧНІ МОДЕЛІ ТА ПРОГРАМНІ ЗАСОБИ РОЗМЕЖУВАННЯ АВТОРСЬКИХ СТИЛІВ АНГЛІЙСЬКОЇ ПРОЗИ

Проаналізовано наявні дослідження щодо встановлення авторства тексту, внаслідок чого з'ясовано, що підвищення достовірності авторської атрибуції тексту є актуальним завданням у контексті тенденції до збільшення загального обсягу текстової інформації в мережі Інтернет. Розроблено модель системи фоностатистичних структур стилів. Достовірність авторської атрибуції підвищено на основі побудованої моделі системи фоностатистичних структур досліджуваних стилів (художнього, розмовного, газетного, публіцистичного, наукового) англійської мови. Складовими компонентами моделі системи фоностатистичних структур досліджуваних стилів є вдосконалені статистичні моделі: модель стильової, підстильової й авторської диференціації текстів за методом гіпотез і ранжування та модель визначення стилерозрізняльної здатності груп приголосних фонем досліджуваних стилів. Перша статистична модель ґрунтується на визначенні ступеня встановлених істотних відмінностей за відношенням кількості груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями до загальної кількості груп приголосних фонем. Істотні розходження визначено за кількістю груп приголосних фонем, за якими встановлено істотні відмінності за різницею значень середніх частот груп приголосних фонем та за різницею значень рангових показників середніх частот груп приголосних фонем. Друга статистична модель ґрунтується на визначенні авторорозрізняльної здатності групи приголосних фонем за відношенням кількості зіставлень, у яких встановлено істотні відмінності між текстами різних авторів до загальної кількості всіх зіставлень. Побудована модель системи фоностатистичних структур досліджуваних стилів англійської мови дала змогу встановити статистичні параметри авторського стилю Е. Бронте на матеріалі твору "Буремний перевал", а також статистичні параметри розмовного, газетного, публіцистичного і наукового стилів. Спрощено процес авторської та стильової атрибуції тексту шляхом зменшення кількості груп приголосних фонем до двох (група передньоязикових і група губних), що забезпечує вищий рівень автоматизації. Вдосконалені статистичні моделі реалізовано на мові програмування Java, що забезпечує платформонезалежність програмного продукту. Структура програми ґрунтується на модульному принципі, що дає змогу швидко модифікувати та вдосконалювати програму.

Ключові слова: модель авторського стилю; фоностатистична структура; метод гіпотез; метод ранжування; стилерозрізняльна здатність групи фонем.

Вступ

Важливість встановлення авторства досліджуваного тексту зростає із збільшенням обсягу текстової інформації в інтернет-просторі. Використовувані методи і методики авторської атрибуції не вирішують завдання повною мірою. Зокрема, актуальним є завдання підвищення достовірності встановлення авторства з використанням моделювання. Моделювання об'єкта дослідження є надійним інструментом для досягнення поставленої дослідником мети, оскільки побудований аналог відображає ті характеристики, особливості та властивості досліджуваного об'єкта, які доцільно докладно вивчити. Для встановлення авторства тексту необхідно виокремити ті параметри, які мають вирізняльний характер, тобто за ними можна відрізнити текст цього автора від тексту іншого автора. Такими параметрами є статистичні параметри авторських стилів. Це частоти, середні частоти та відносні частоти. Для статистичного моделювання дослідники використовують різні статистичні методи. Зокрема, це регресійний аналіз та інші методи машинного навчання [1, 7, 8, 12, 14, 15], критерій Ст'юдента [9, 10].

Інформація про авторів:
Хомицька Ірина Юріївна, асистент, кафедра прикладної лінгвістики.
Email: iryna.khomytska@ukr.net; https://orcid.org/0000-0003-3470-7197
Теслюк Василь Миколайович, д-р техн. наук, професор, завідувач кафедри автоматизованих систем управління.
Email: vasyl.m.teslyuk@lpnu.ua; https://orcid.org/0000-0002-5974-9310
Базилевич Ірина Богданівна, канд. фіз.-мат. наук, доцент, кафедра теоретичної та прикладної статистики.
Email: i_bazylevych@yahoo.com
Береговський Василь Васильович, канд. техн. наук, доцент, кафедра комп'ютерних систем та мереж.
Email: beregovskyivasyl@gmail.com

Цитування за ДСТУ: Хомицька І. Ю., Теслюк В. М., Базилевич І. Б., Береговський В. В. Статистичні моделі та програмні засоби розмежування авторських стилів англійської прози. Науковий вісник НЛТУ України. 2020, т. 30, № 5. С. 135–139.
Citation APA: Khomytska, I. Yu., Teslyuk, V. M., Bazylevych, I. B., & Beregovskyi, V. V. (2020). The statistical models and software for authorial style differentiation in english prose. *Scientific Bulletin of UNFU*, *30*(5), 135–139. https://doi.org/10.36930/40300522

Аналіз останніх досліджень та публікацій. З-поміж основних чинників, які впливають на достовірність результатів авторської та стильової атрибуції потрібно виокремити рівень мови, на якому розмежовуються тексти, ефективність використовуваних методів на вибраному рівні мови, доречність застосування поєднання методів та визначення загальної стильової маркованості для зіставлюваних текстів різних авторів. У більшості останніх публікацій відстежується тенденція розмежовувати тексти на лексико-семантичному рівні. У цьому напрямі розрізняють лексеми із високою частотою вживання, наявні у переважній більшості текстів, лексеми з середньою частою вживання та характерні для авторського стилю лексеми з низькою частотою вживання [8]. Використовується методика визначення частоти вживання лексем завдовжки 2-5 букв, знаків пунктуації, пробілів тощо [15]. У цьому зв'язку, ефективним є комплексний підхід, який дає змогу визначити оптимальне поєднання мовних одиниць та символів для визначення характеристик авторського стилю [14]. Поєднання лексем із середньою та низькою частотою вживання забезпечує кращі результати, ніж використання лексем з високою частотами вживання [7]. Моделювання граматичних категорій зумовлює застосування конкретних методів і методик породжувальних граматик [12]. Моделювання лексико-семантичного пласту мови є доречним для визначення як змісту тексту, так і характерних авторських особливостей [1]. Використання статистичних структур для моделювання авторських стилів дає змогу розкрити специфіку авторської манери викладу [9, 10]. Зазначивши позитивні аспекти наведених вище досліджень, варто зауважити, що актуальним є підвищення достовірності авторської та стильової атрибуції, яке у цьому дослідженні забезпечується вибором фонологічного рівня з найстрогішою структурою для формалізації розмежування стилів та побудовою моделі системи фоностатистичних структур стилів, яка визначає ступінь встановлених істотних відмінностей між текстами різних авторів.

Об'єкт дослідження – процес атрибуції англомовних текстів з використанням статистичних моделей.

Предмет дослідження – статистичні моделі та програмні засоби автоматизації процесу атрибуції англомовних текстів, що дає змогу підвищити рівень достовірності та автоматизації встановлення авторства.

Мета роботи – підвищення рівня достовірності та автоматизації авторської атрибуції англомовних текстів, що забезпечить вищу ефективність встановлення авторства.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження:*

- вдосконалити статистичну модель стильової, підстильової та авторської диференціації текстів;
- вдосконалити статистичну модель визначення стилерозрізняльної здатності груп приголосних фонем.

Наукова новизна отриманих результатів дослідження – вперше побудовано модель системи фоностатистичних структур досліджуваних стилів англійської мови, яка є аналогом системи функціональних стилів. Складовими частинами цієї моделі системи є вдосконалені статистичні моделі: модель стильової, підстильової та авторської диференціації текстів за методом гіпотез і ранжування та модель визначення стилерозрізняльної здатності груп приголосних фонем досліджуваних стилів.

Практична значущість результатів дослідження – результати дослідження можна використовувати у тих сферах людської діяльності, де необхідно встановити авторство тексту. Це сфера судочинства (встановлення авторства заповітів, показів свідків, правових документів), сфера науки (встановлення авторства наукових статей), сфера художньої літератури (встановлення авторства художніх творів).

Результати дослідження та їх обговорення

Об'єктом статистичного моделювання у цьому дослідженні є статистична сукупність авторського стилю Емілі Бронте, в якій реалізується закономірність розподілу частот груп приголосних фонем, яка є вирізняльною характеристикою індивідуальної манери викладу письменниці [4].

Побудовані моделі є спрощеним, математично-формалізованим способом апроксимації авторського стилю Емілі Бронте. На основі даних, отриманих із спостережень з вибірок із твору "Буремний перевал" (розділи 17-34), обчислюються статистичні відношення між частотами груп приголосних фонем, використовуючи статистичний критерій Колмогорова-Смірнова, критерій χ-квадрат та критерій Ст'юдента [4]. Зазначене поєднання статистичних критеріїв дає змогу підвищити достовірність авторської атрибуції, а використання однієї з восьми груп приголосних фонем забезпечує вищий рівень автоматизації розмежування текстів різних авторів. Практична значущість дослідження передбачає використання отриманих результатів у судочинстві для встановлення автора судових документів, у сфері науки для ідентифікації автора наукових статей і загалом, у всіх сферах діяльності, в яких необхідно встановити автора тексту.

1. ***Моделі диференціації фоностатистичних структур стилів***. У дослідженні виконується завдання встановлення авторства англомовних текстів. У математичній статистиці такі задачі розв'язують шляхом перевірки з використанням гіпотези на однорідність [2, 11, 17]. Суть цієї гіпотези полягає в тому, що необхідно перевірити: чи отримані дві вибірки є вибірками з одного розподілу. Доречно навести статистичний опис задачі. Нехай досліджується деяка випадкова величина X. Для цього здійснюється серія спостережень, або серія стохастичних експериментів. Результат першого спостереження – це випадкова величина, яка позначається X_1, результат другого спостереження – X_2, …, результат n-го спостереження – X_n. Сукупність усіх спостережень – це випадковий вектор $(X_1,...,X_n)$, який позначається $\tilde{X}$. Тобто $\tilde{X}=(X_1,...,X_n)$. Отже, вибірка – це випадковий вектор. Вважаємо, що результати спостережень є незалежними і однаково розподіленими.

Для конкретного дослідження розглядається не випадковий вектор, а числовий вектор, бо статистичними даними є певні числові значення. Ці числові значення позначаються $x_1,...,x_n$, їх сукупність – числовий вектор $\tilde{x}=(x_1,...,x_n)$. Вектор $\tilde{x}$ є реалізацією вибірки.

Зіставляються два тексти. Обсяг кожної вибірки становить 51000 фонем. Вибірка ділиться на 51 порцію. Підраховується кількість фонем для кожної із восьми груп фонем (губних, передньоязикових, дорсальних, задньоязикових, сонорних, носових, щілинних, зімкне-

них) для кожної порції (1000 фонем). Унаслідок отримується вісім вибірок обсягом у 51 тисячу фонем для кожного тексту. Отже, вибірка X – це вектор, складений із кількості фонем певної групи на кожну порцію у першому тексті, а вибірка Y – це вектор, складений із кількості фонем цієї ж групи на кожну порцію у другому тексті. Отже, необхідно порівняти по дві реалізації вибірки вісім разів, тобто для кожної групи фонем, маємо дві реалізації вибірок з двох текстів. Кожен елемент реалізації вибірки може набувати одне з 1001 значень, тобто від 0 до 1000, тому є сенс вважати цю вибірку – вибіркою з неперервного розподілу. Спочатку перевіряється гіпотеза на вид розподілу, де за гіпотетичний розподіл береться нормальний розподіл. Для всіх груп фонем гіпотеза не відхиляється. Тому, при перевірці на однорідність можна використовувати критерій Ст'юдента [3, 13]. Для восьми груп приголосних фонем у двох текстах висувається вісім нульових гіпотез H_0 – отримані вибірки є вибірками з одного розподілу.

Очевидно, що під час перевірки гіпотези на однорідність доцільно використовувати декілька критеріїв. Використовують такі критерії: критерій Колмогорова-Смірнова, критерій χ-квадрат та критерій Ст'юдента.

У дослідженні вдосконалено статистичну модель стильової, підстильової та авторської диференціації текстів за методами гіпотез і ранжування, яка на відміну від наявних, враховує позицію фонеми у слові, визначає відношення кількості груп приголосних фонем, за якими встановлено істотні відмінності, між попарно зіставленими стилями до загальної кількості груп приголосних фонем. Ідею розмежування стилів за кількістю груп приголосних фонем – пропонуємо вперше. Велика кількість груп приголосних фонем (6-8), за якими тексти відрізняються істотно, дає змогу з більшою достовірністю робити висновки про їх відмінність.

Вдосконалена статистична модель стильової, підстильової та авторської диференціації текстів за методом гіпотез і ранжування реалізується таким алгоритмом:

Крок 1. Встановити істотні відмінності для восьми груп приголосних фонем у зіставленні стилів, підстилів і текстів різних авторів, перевіряючи статистичні гіпотези з використанням критерію Ст'юдента [2, 3, 4, 11, 17], критерію Колмогорова-Смірнова [6, 16] та критерію χ-квадрат [17] з довірчою ймовірністю 0,95.

Крок 2. Розмежувати стилі, підстилі та тексти різних авторів за встановленими істотними відмінностями, використовуючи метод ранжування.

Крок 3. Підрахувати кількість груп приголосних фонем, за якими встановлено істотні відмінності.

Крок 4. Визначити ступінь встановлених істотних відмінностей (DSD) за відношенням кількості груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями (*SDGN*) до загальної кількості груп приголосних фонем (*TGN*): $DSD = SDGN / TGN$.

Крок 5. Встановити низький ступінь встановлених істотних відмінностей (*LDSD*) за відношенням 1-3 груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями до загальної кількості груп приголосних фонем – 8: $LDSN = 1/8 = 0{,}13$; $LDSN = 2/8 = 0{,}25$; $LDSN = 3/8 = 0{,}38$.

Крок 6. Встановити середній ступінь встановлених істотних відмінностей (*MDSD*) за відношенням 4-5 груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями до загальної кількості груп приголосних фонем – 8: $MDSN = 4/8 = 0{,}5$; $MDSN = 5/8 = 0{,}63$.

Крок 7. Встановити високий ступінь встановлених істотних відмінностей (*HDSD*) за відношенням 6-8 груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями до загальної кількості груп приголосних фонем – 8: $HDSN = 6/8 = 0{,}75$, $HDSN = 7/8 = 0{,}88$, $HDSN = 8/8 = 1$.

Отже, вдосконалено статистичну модель визначення стилерозрізняльної здатності груп приголосних фонем, яка базується на визначенні відношення кількості порівнянь, за якими виявлено істотні відмінності між стилями – до кількості всіх порівнянь. Зменшення кількості груп приголосних фонем, за якими диференціюються стилі, спрощує процес атрибуції тексту, підвищує його достовірність і забезпечує вищий рівень автоматизації шляхом здійснення атрибуції тексту за групою приголосних фонем з найбільшою стилерозрізняльною здатністю. Вдосконалена статистична модель визначення стилерозрізняльної здатності груп приголосних фонем досліджуваних стилів реалізується таким алгоритмом:

Крок 1. Розмежувати тексти за критерієм Ст'юдента, Колмогорова-Смірнова та критерієм χ-квадрат.

Крок 2. Визначити групу фонем, за якою встановлено істотні відмінності за критерієм Ст'юдента, Колмогорова-Смірнова та критерієм χ-квадрат: $ADC_1 = (t + \lambda_{n,m} + \hat{\chi}^2_n)/3$, де ADC_1 – здатність групи фонем розрізняти авторські стилі; t – значення критерію Ст'юдента, $\lambda_{n,m}$ – значення критерію Колмогорова-Смірнова, $\hat{\chi}^2_n$ – значення критерію χ-квадрат.

Крок 3. Визначити групу фонем, за якою встановлено істотні відмінності у всіх випадках позиції фонеми у слові: $ADC_2 = (UP + IP + FP)/3$, де ADC_2 – здатність групи фонем розрізняти авторські стилі у трьох випадках позиції фонеми у слові: *UP* – невизначена позиція фонеми у слові, *IP* – фонема на початку слова, *FP* – фонема у кінці слова.

Крок 4. Визначити авторорозрізняльну здатність групи приголосних фонем (*ADC*) за відношенням кількості зіставлень, у яких встановлено істотні відмінності між текстами різних авторів (*SSD*) до загальної кількості всіх зіставлень (*TNC*): $ADC = SSD / TNC$.

Крок 5. Визначити групу фонем з найбільшим значенням *ADC*.

Крок 6. Розмежувати тексти за цією групою фонем.

Отже, вдосконалені статистичні моделі дають змогу підвищити достовірність результатів авторської атрибуції на основі застосування трьох статистичних критеріїв – критерію Колмогорова-Смірнова (К-С), критерію χ-квадрат (Х-К) та критерію Ст'юдента (Ст.) та підвищити рівень автоматизації, використовуючи одну замість восьми групу приголосних фонем.

2. *Особливості реалізації програмної системи та результати дослідження*. Вдосконалені статистичні моделі реалізуються на мові програмування Java. Розроблена програмна система автоматизації диференціації фоностатистичних структур функціональних стилів англійської мови ґрунтується на модульному принципі. Структура програмної системи складається із шести модулів: модуля введення/виведення даних, модуля пе-

ретворення англомовного тексту в транскрипційний варіант, модуля формування вибірки з приголосних фонем, модуля обчислення кількості фонем у вибірці, модуля визначення середньої та відносної частоти вживання груп приголосних фонем та модуля статистичного опрацювання вибірки використовуючи критерій К-С, критерій Х-К та критерій Ст.

Розроблено алгоритм функціонування програми відповідно до вдосконаленої статистичної моделі стильової, підстильової та авторської диференціації текстів за методом гіпотез і ранжування – Модель 1 (рис. 1).

Розроблено алгоритм функціонування програми відповідно до вдосконаленої статистичної моделі визначення стилерозрізняльної здатності груп приголосних фонем досліджуваних стилів – Модель 2 (рис. 2).

Рис. 1. Блок-схема алгоритму функціонування програми відповідно до Моделі 1

Структура розробленого програмного забезпечення складається з п'яти класів: класу Main, який забезпечує послідовність виконання дій у програмі, класу Transcription Processor, який відповідає за перетворення букв у транскрипційні символи, класу ConsonantProcessor, який дає змогу створити вибірку з приголосних фонем, класу ConsonantUtils, який забезпечує обчислення кількості фонем по групах і порціях, класу StatisticProcessor, який відповідає за розмежування вибірок за К-С, Х-К. Ст.

Розроблене інформаційне забезпечення ґрунтується на використанні спискових структур ArrayList<String>, які є зручними у роботі з динамічними даними. Транскрипційні варіанти текстів зберігається у структурі даних HashMap, яка побудована на принципі ключ-значення і дає змогу уникнути дублікатів.

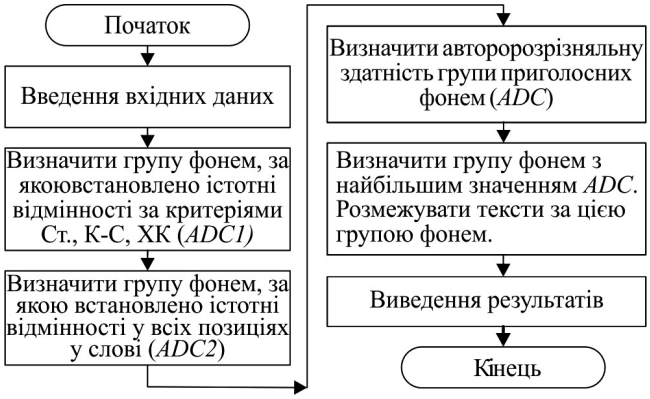

Рис. 2. Блок-схема алгоритму функціонування програми відповідно до Моделі 2

Програмна реалізація вдосконалених статистичних моделей дала змогу отримати статистичні параметри авторського стилю Емілі Бронте з вибірок із твору "Буремний перевал" (розділи 17-34). Статистичні параметри авторського стилю Емілі Бронте, визначені з вибірок попередніх розділів (розділи 1-16), наведено у попередній роботі [5]. Для розділів 17-34 істотні відмінності встановлено за чотирма групами приголосних фонем: групами велярних, носових, передньоязикових і губних. Найвищу авторорозрізняльну здатність мають тільки дві групи приголосних фонем – передньоязикові і губні, бо за цими групами фонем встановлено істотні відмінності, використовуючи три статистичні критерії – критерій К-С, критерій Х-К та критерій Ст. Результати обчислення наведено у таблиці.

Таблиця. Результати визначення авторорозрізняльної здатності груп фонем

Група фонем	Критерій Ст'юдента	Критерій Колмогорова-Смірнова	Критерій χ-квадрат
велярні		+	
щілинні			
носові	+		
сонорні			
передньоязикові	+	+	+
дорсальні			
зімкнені			
губні	+	+	+

Модель визначення авторорозрізняльної здатності груп приголосних фонем дала змогу встановити істотні відмінності за групою передньоязикових фонем у зіставленні вибірок з 17-34 розділів твору Е. Бронте "Буремний перевал". Структурна схема на рис. 3,*а* відображає результати досліджень.

Результати розмежування вибірок з художньої прози Е. Бронте за групою губних фонем наведено на рис. 3,*б*.

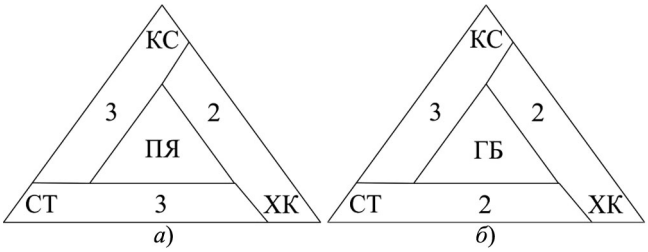

Рис. 3. Структурна схема подання результатів за групою передньоязикових фонем (*а*) та групою губних фонем (*б*): КС – критерій Колмогорова-Смірнова, ХК – критерій χ-квадрат та СТ – критерій Ст'юдента; ПЯ – група передньоязикових фонем; ГБ – група передньоязикових фонем; (2, 3, 3), (2, 2, 3) – кількість зіставлень, у яких встановлено істотні відмінності

Обговорення результатів дослідження. Розроблена модель системи фоностатистичних структур досліджуваних стилів, яка містить дві вдосконалені статистичні моделі, є новизною дослідження. На відміну від попередніх робіт інших дослідників цього безпосередньо [1, 7, 8, 9, 10, 12, 14, 15], розроблена модель забезпечує підвищення достовірності авторської атрибуції шляхом визначення ступенів істотних відмінностей зіставлених текстів художньої прози Е. Бронте та текстів художнього, розмовного, газетного, публіцистичного і наукового стилів англійської мови.

Висновки

Розроблено модель системи фоностатистичних структур досліджуваних стилів (художнього, розмовного, газетного, публіцистичного, наукового) англійської мови. Модель складається із двох вдосконалених статистичних моделей. Вдосконалено статистичну модель сти-

льової, підстильової та авторської диференціації текстів за методами гіпотез і ранжування, яка визначає відношення кількості груп приголосних фонем, за якими встановлено істотні відмінності між попарно зіставленими стилями до загальної кількості груп приголосних фонем. Статистичні параметри визначено для розділів 17-34 твору Е. Бронте "Буремний перевал".

Вдосконалено статистичну модель визначення стилерозрізняльної здатності груп приголосних фонем, яка базується на визначенні відношення кількості порівнянь, за якими виявлено істотні відмінності між стилями – до кількості всіх порівнянь. Зменшення кількості груп приголосних фонем, за якими диференціюються стилі до двох, спрощує процес атрибуції тексту, підвищує його достовірність і забезпечує вищий рівень автоматизації авторської атрибуції тексту. У дослідженні вибірки з художньої прози Е. Бронте розмежовано за групами передньоязикових і губних фонем.

Розроблено програмні засоби, які дають змогу з більшою достовірністю здійснити стильову та авторську атрибуцію тексту. Наведено результати, які дають змогу ствердити, що група передньоязикових і губних фонем має найвищу авторорозрізняльну здатність у зіставленні досліджуваних текстів.

References

1. Davydov, M., & Lozynska, O. (2016). Linguistic Models of Assistive Computer Technologies for Cognition and Communication. *Proceedings of the XIth Scientific and Technical Conference*, (CSIT'2016), Lviv, Ukraine, 171–174.
2. Hnedenko, B. V. (2010). *Kurs teorii ymovirnostei*. Kyiv: Kyivskyi universytet, 464 p. [In Ukrainian].
3. Jones, M. C. (2002). Students simplest distribution. *Journal of the Royal Statistical Society*. Series D, 51, 41–49.
4. Khomytska, I., Teslyuk, V., Holovatyy, A., & Morushko, O. (2018). Development of Methods, Models and Means for the Author Attribution of a Text. Eastern-European *Journal of Enterprise Technologies, 3/2*(93), 41–46. https://doi.org/10.15587/1729-4061.2018.132052
5. Khomytska, I., Teslyuk, V., Kryvinska, N., & Bazylevych, I. (2020, July). Software-Based Approach Towards Automated Authorship Acknowledgement – Chi-Square Test on One Consonant Group. *Electronics, 4*(7), 1138. https://doi.org/10.3390/electronics9071138
6. Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Chelsea Publishing, 340 p.
7. Koppel, M., Schler, J., & Argamon, Sh. (2011). Authorship attribution in the wild. *Language Resources and Evaluation, 45*(1), 46–52. https://doi.org/10.1007/s10579-009-9111-2
8. Madigan, D., Genkin, A., Lewis, D. D, Argamon, Sh., Fradkin, D.,& Li, Ye. (2005). Author Identification on the Large Scale. *AIP Conference Proceedings 803*, 509–5013. https://doi.org/10.1063/1.2149832
9. Perebyinis, V. S. (1967). *Statystychni parametry styliv*. Kyiv: Scientific thought, 240 p. [In Ukrainian].
10. Perebyinis, V. S. (2013). *Statystychni metody dlia linhvistiv*. Vinnytsia: Nova Knyha, 170 p. [In Ukrainian].
11. Seno, P. S. (2004). *Teoriia ymovirnostei ta matematychnoi statystyky*: pidruchnyk. Kyiv: Tsentr navchalnoi literatury, 448 p. [In Ukrainian].
12. Shestakevych, T., Vysotska, V., Chyrun, L., & Chyrun, L. (2014). Modelling of semantics of natural language sentences using generative grammars. *Computer Science and Information Technologies: Proceedings of the IX-th Int. Conference*, (CSIT'2014), 18–22 November, 2014, Lviv, Ukraine, 19–22.
13. Snedecor, G. W., & Cochran, W. G. (1989). *Statistical Methods*. Iowa; Iowa State Press. USA, 438 p.
14. Stamatatos, E. (2017). Authorship attribution using text distortion. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, (Vol. 1), pp. 1138–1149.
15. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., Lopez Lopez, A., Potthast, M., & Stein, B. (2015). Overview of the Author Identification Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings. CLEF and CEUR-WS.org*.
16. Steinskog, D. J. (2007). *A cautionary note on the use of the Kolmogorov-Smirnov test for normality*. American Meteor Soc., *135*, 1151–1157.
17. Turchyn, V. M. (2014). *Teoriia ymovirnostei i matematychna statystyka*. Dnipropetrovsk: IMA-pres, 294 p. [In Ukrainian].

I. Yu. Khomytska[1], V. M. Teslyuk[1], I. B. Bazylevych[2], V. V. Beregovskyi[3]

[1] Lviv Polytechnic National University, Lviv, Ukraine
[2] Ivan Franko National University of Lviv, Lviv, Ukraine
[3] Ivano-Frankivsk National Technical University of Oil and Gas, Ivano-Frankivsk, Ukraine

THE STATISTICAL MODELS AND SOFTWARE FOR AUTHORIAL STYLE DIFFERENTIATION IN ENGLISH PROSE

The conducted analysis of authorship attribution has shown that the problem of test validity enhancement is a topical problem, as the general amount of text information on the Internet is constantly increasing. To solve this problem, a model of the system of phonostatistical structures of the researched styles (belles-lettres, colloquial, newspaper, publicist, scientific) of the English language has been built. The constituents of the model of the system of phonostatistical structures of the researched styles are a statistical model of style and authorial differentiation by the hypothesis and ranking methods and a statistical model of determination of style-differentiating capability of a consonant phoneme group. The first statistical model is based on defining the degree of statistically significant differences between the compared texts by the number of consonant phoneme groups in which the styles differ essentially to the general number of consonant phoneme groups. The significant differences have been determined by the number of consonant phoneme groups in which the essential differences have been established by the difference of values of the mean frequencies of occurrence of consonant phoneme groups and the difference of values of ranking indices of the mean frequencies of occurrence of consonant phoneme groups. The second statistical model is based on the determination of the authorial style-differentiating capability of a consonant phoneme group by a relation of the number of comparisons in which the texts by different authors and the general number of all comparisons differ essentially. The improved statistical models have enabled enhancing the test validity of authorship and style attribution as well as establishing the statistical parameters for the texts by E. Bronte and the texts from the belles-lettres, colloquial, newspaper, publicist and scientific functional styles. The authorship attribution performed in one consonant phoneme group instead of eight has considerably simplified the whole process. The models have been implemented on the Java programming language. The software is based on a structure of modules which allows quick modifying and improving the program.

Keywords: model of an author's writing style; phonostatistical structures; hypothesis method; ranking method; style-differentiating capability of a consonant phoneme group.