Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
http://nv.nltu.edu.ua
https://doi.org/10.15421/40290731
Article received 27.06.2019 p.
Article accepted 26.09.2019 p.
УДК 519.765

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], В. В. Береговський[2]

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

МАТЕМАТИЧНІ МЕТОД І МОДЕЛЬ ДИФЕРЕНЦІАЦІЇ ФОНОСТАТИСТИЧНИХ СТРУКТУР АВТОРСЬКОГО СТИЛЮ

Розроблено метод комплексного аналізу диференціації фоностатистичних структур авторського стилю англійської мови, який ґрунтується на поєднанні трьох статистичних критеріїв: критерію Стьюдента, критерію Колмогорова-Смірнова і критерію хі-квадрат. Поєднання цих критеріїв дає змогу підвищити достовірність диференціації авторських стилів. Для розв'язання задачі диференціації авторських стилів побудовано статистичну модель, яка підвищує достовірність результатів авторської атрибуції тексту. Розроблена програмна система реалізує метод і модель з використанням мови програмування Java, що забезпечує платформонезалежність. Для тестування програми вибрано статті С. Логан і Д. Вебстер з газети "Вільна газета" ("Freedom Paper", papers by S. Logan and D. Webster). Істотні відмінності встановлено за групами носових, дорсальних і велярних фонем за критерієм Стьюдента, за всіма вісьмома групами фонем за критерієм Колмогорова-Смірнова, істотні відмінності встановлено за групами сонорних, щілинних, дорсальних, зімкнених, губних і носових фонем за критерієм хі-квадрат. Поєднання використаних критеріїв дало змогу встановити групу фонем з найбільшою авторорозрізняльною здатністю. Це група дорсальних фонем. За цією групою можна диференціювати тексти різних авторів, що дає змогу змінімізувати кількість груп фонем, за якими здійснюється авторська атрибуція тексту.

Ключові слова: фоностатистична структура стилю; диференціація текстів; авторська атрибуція тексту.

Вступ. Однією з основних тенденцій розвитку сучасного суспільства є глобалізація з використанням засобів Інтернету. Щоденно опрацьовуються великі обсяги інформації. У випадках, коли автор тексту невідомий, потрібно встановити його авторство. Тому актуальною задачею сьогодення є авторська атрибуція тексту. Встановлення авторства тексту ґрунтується на диференціації текстів. Оскільки мова є імовірнісною системою, тоді, відповідно, тексти потрібно диференціювати імовірнісними методами. Аналіз сучасного стану досліджень, в яких застосовуються методи математичної статистики, показав, що найчастіше тексти диференціюються на фонологічному, лексико-семантичному і синтаксичному рівнях мови. Так, на фонологічному рівні тексти диференціюються вітчизняними дослідниками – В. Перебийніс, В. Левицьким (Perebyinis, 2013; Levitckii, 2007) та зарубіжними дослідниками – Д. Сегалом, Р. Піотровським (Segal, 1968, 1972; Piotrovskii, 1999); на лексико-семантичному рівні, вітчизняними дослідниками – В. Перебийніс, В. Левицьким, В. Литвином, В. Висоцькою (Lytvyn et al., 2017a, 2017b) та зарубіжними дослідниками – Г. Альтманом, Р. Колером, Ш. Аргамоном, М. Копелем, Дж. Шлером, М. Пенбейкером (Altman, 2005; Altmann, Kohler & Piotrowski, 2005; Argamon et al., 2009; Koppel, 2009); на синтаксичному рівні, вітчизняними дослідниками – В. Перебийніс, В. Левицьким, О. Бісікало (Bisikalo & Vysotska, 2016). У цьому дослідженні вибрано фонологічний рівень, оскільки він має незмінну кількість елементів, строгішу структуру і його легше формалізувати. Аналіз робіт зазначених дослідників показав, що для диференціації двох текстів використовується один статистичний метод. Однак при диференціації двох тих самих текстів використання різних статистичних методів дає різні результати. Тому актуальним є розроблення методу комплексного аналізу, який ґрунтується на поєднанні статистичних методів і дає змогу підвищити достовірність диференціації текстів. Відповідно, метою дослідження є розроблення методу та моделі, які дають змогу підвищити достовірність авторської атрибуції текстів. Для досягнення поставленої мети потрібно розробити метод комплексного аналізу та статистичну модель диференціації фоностатистичних структур стилів та інструментальні засоби для автоматизації процесу авторської атрибуції англомовних текстів.

Методи та моделі диференціації фоностатистичних структур авторського стилю. У дослідженні тексти різних авторів продиференційовано розробленим

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, кафедра систем автоматизованого проектування.
Email: vasyl.m.teslyuk@lpnu.ua; https://orcid.org/0000-0002-5974-9310
Береговський Василь Васильович, канд. техн. наук, доцент, кафедра комп'ютерних систем і мереж.
Email: beregovskyivasyl@gmail.com
Цитування за ДСТУ: Хомицька І. Ю., Теслюк В. М., Береговський В. В. Математичні метод і модель диференціації фоностатистичних структур авторського стилю. Науковий вісник НЛТУ України. 2019, т. 29, № 7. С. 156–159.
Citation APA: Khomytska, I. Yu., Teslyuk, V. M., & Beregovskyi, V. V. (2019). Mathematical Method and Model of Differentiation of Phonostatistical Structures of Authorial Style. *Scientific Bulletin of UNFU, 29*(7), 156–159. https://doi.org/10.15421/40290731

методом комплексного аналізу диференціації фоностатистичних структур стилів, який ґрунтується на поєднанні критеріїв методу гіпотез (критеріїв Стьюдента, Колмогорова-Смірнова і хі-квадрат). Розроблений алгоритм методу комплексного аналізу диференціації фоностатистичних структур стилів складається з таких кроків:

Крок 1. Введення вхідних даних.
Крок 2. Перевірка, чи вибірки є з нормального розподілу.
Крок 3. Диференціація текстів за критерієм Стьюдента.
Крок 4. Диференціація текстів за критерієм Колмогорова-Смірнова.
Крок 5. Диференціація текстів за критерієм хі-квадрат.
Крок 6. Аналіз результатів, отриманих за трьома критеріями.
Крок 7. Виведення результатів.

Відповідно до розробленого алгоритму реалізації методу, початкові вибірки перевірено на нормальність за критерієм Пірсона (Khomytska, Teslyuk & Shakhovska, 2016; Gries, 2009; Khomytska et al., 2018); продиференційовано за критерієм Стьюдента:

$$t = \frac{\bar{x}_1^a - \bar{x}_2^a}{S}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

де $\bar{x}_1^a - \bar{x}_2^a$ – різниця середніх частот 1-ї і 2-ї вибірок за фіксованою групою фонем; n_1 і n_2 – кількість порцій 1-ї і 2-ї вибірок; продиференційовано за критерієм Колмогорова-Смірнова:

$$\lambda_{n,m} = \sqrt{\frac{nm}{n+m}} D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{-\infty<z<\infty} \left|F_n(z) - F_m(z)\right|,$$

де $D_{n,m} = \sup_{-\infty<z<\infty} \left|F_n(z) - F_m(z)\right|$; $F_n(z)$ і $F_m(z)$ – емпіричні функції розподілу, побудовані для двох вибірок n і m, $\lambda_{n,m}$ – значення статистики Колмогорова-Смірнова; продиференційовано за критерієм Колмогорова-Смірнова:

$$\hat{\chi}_n^2 = \sum_{i=1}^{s}\sum_{j=1}^{k} \frac{\left(\upsilon_{i,j} - \frac{n_j \upsilon_j}{n}\right)^2}{n_j \upsilon_j \cdot n}, \quad \upsilon_j = \sum_{j=1}^{k} \upsilon_{ij},$$

де: $\upsilon_{i,j}$ – кількість елементів, які потрапили в i-й інтервал j-го тексту; s – кількість груп фонем; k – кількість текстів; n – кількість порцій у двох текстах; n_j – кількість порцій в одному тексті.

На підставі побудованого методу комплексного аналізу диференціації фоностатистичних структур стилів, розроблено модель визначення авторо-розрізняльної здатності груп приголосних фонем (АЗФ), яка базується на виразі:

$$АЗФ = \frac{t + \lambda_{n,m} + \chi^2}{3},$$

де: t – значення критерію Стьюдента; $\lambda_{n,m}$ – значення критерію Колмогорова-Смірнова; χ_n^2 – значення критерію хі-квадрат. Модель дає змогу встановити групу фонем з найбільшою авторорозрізняльною здатністю, і цим самим мінімізувати кількість груп фонем, за якими здійснюється авторська атрибуція тексту.

Особливості реалізації програмної системи та результати дослідження. Розроблене математичне забезпечення реалізовано у програмній системі, структура якої включає такі модулі: модуль введення/виведення даних; модуль перетворення тексту в транскрипційний варіант; модуль визначення кількості приголосних фонем у вибірці; модуль визначення середніх частот груп приголосних фонем; модуль перевірки тексту на нормальність (критерій Пірсона); модуль визначення критеріїв Стьюдента, Колмогорова-Смірнова і хі-квадрат (Juala, 2008; Khomytska, Teslyuk& Shakhovska, 2018; Khomytska, Tesliuk & Labinska, 2018).

Розроблене програмне забезпечення (ПЗ) системи диференціації фоностатистичних структур стилів реалізовано з використанням таких класів (рис. 1).

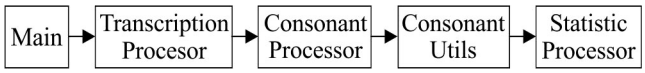

Рис. 1. Структура класів ПЗ програми диференціації фоностатистичних структур стилів

Клас Main призначений для завантаження файлів; за допомогою класу TranscriptionProcessor англомовний текст перетворюється в його транскрипційний варіант; клас ConsonantProcessor призначений для формування вибірок з приголосних фонем. Підрахунок кількості приголосних фонем у кожній порції і об'єднання їх у групи реалізується за допомогою класу ConsonantUtils. У процесі обчислення використовується структура даних List <Map <ConsonantType, Long", яку зображено на рис. 2.

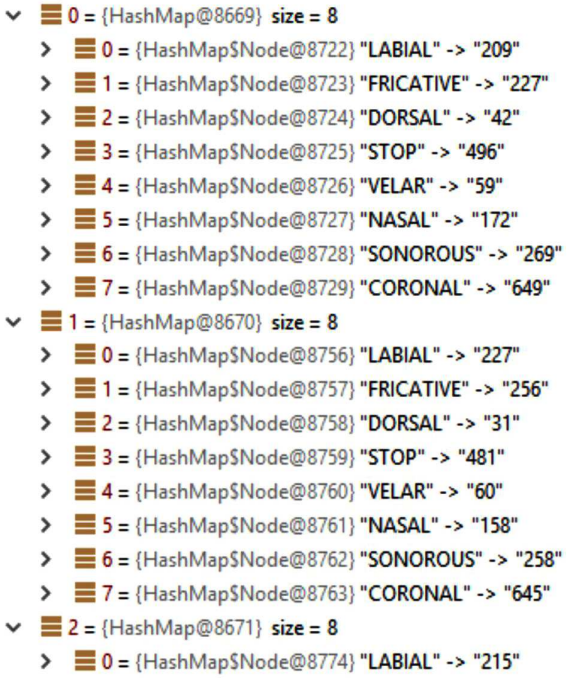

Рис. 2. Структура даних: List<Map<ConsonantType, Long"

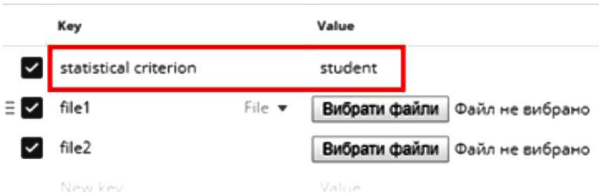

Рис. 3. Меню вибору критерію перевірки гіпотези на однорідність вибірки

Наступним кроком реалізації методу є перевірка двох вибірок на однорідність. Якщо вибірки неоднорідні, то вони відрізняються авторським стилем. Програма перевіряє вибірки на однорідність за критеріями Стьюдента, χ^2 і Колмогорова-Смірнова. Для вибору критерію потрібно у полі "statistical criterion" зазначити його назву (рис. 3). Клас StatisticProcessor розробленого ПЗ, представлений на UML діаграмі, відповідає за статистичне опрацювання вибірок (рис. 4).

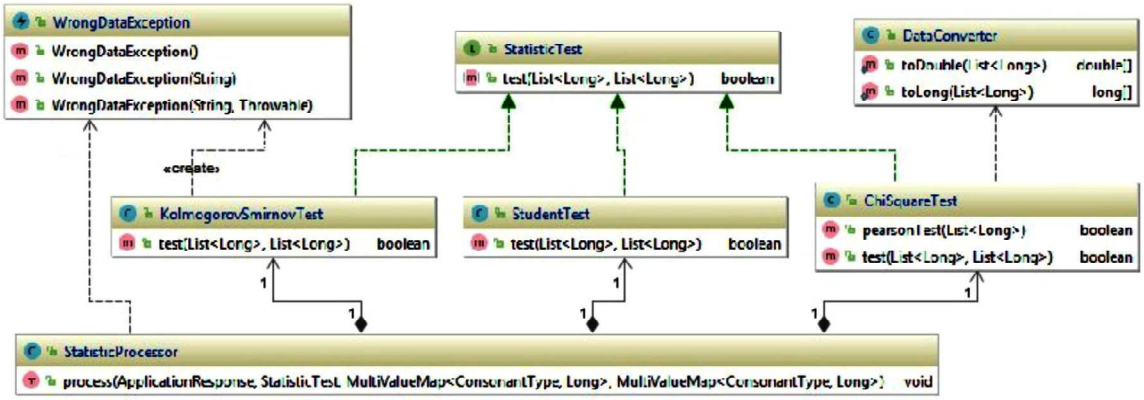

Рис. 4. Структура класів, які відповідають за статистичне опрацювання вибірки

Інформаційне забезпечення системи диференціації фоностатистичних структур функціональних стилів включає вбудовану базу даних H2, написану мовою програмування Java.

Із завантаженням нових текстів відбувається динамічна зміна транскрипційних знаків. Для збереження даних використано списокову структуру даних ArrayList<String>, яка є автоматичним розширювальним списком, і за допомогою якої зручно виконувати операції з динамічними даними. Приклад списокової структури даних зображено на рис. 5.

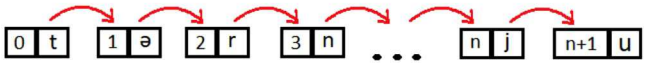

Рис. 5. Приклад списокової структури даних

У процесі застосування методу кожне слово нового тексту зіставляється з уже наявними словами у програмі. Із збігом таких слів, у програму передається команда, що слово з індексом "x" потребує транскрипційного варіанта. Список видає стрічку з перетворенням у транскрибційні знаки з відповідним індексом. Якщо слова немає, тоді до списокової структури записується символ з цим самим індексом, що і його перетворення у списку. Статтю С. Логан і Д. Вебстер з газети "Вільна газета" ("Freedom Paper", papers by S. Logan and D. Webster) було вибрано для тестування функціонування програми. Встановлено, що істотні відмінності за групами носових, дорсальних і велярних фонем за критерієм Стьюдента, за всіма вісьмома групами фонем за критерієм Колмогорова-Смірнова, істотні відмінності встановлено за групами сонорних, щілинних, дорсальних, зімкнених, губних і носових фонем за критерієм хі-квадрат (рис. 6).

```
24      "Statistic Criterion": "CHI_SQUARE",
25      "Statistic Results": {
26          "SONOROUS": false,
27          "FRICATIVE": false,
28          "CORONAL": true,
29          "DORSAL": false,
30          "STOP": false,
31          "LABIAL": false,
32          "NASAL": false,
33          "VELAR": true
34      },
35      "Error message": []
```

Рис. 6. Результати дослідження з використанням критерію хі-квадрат

Для розв'язання задачі визначення авторорозрізняльної здатності групи велярних фонем побудовано статистичну модель за результатами зіставлення текстів публічних промов американського президента Д. Трампа і статей С. Логан та Д. Вебстер з газети "Вільна газета" (рис. 7).

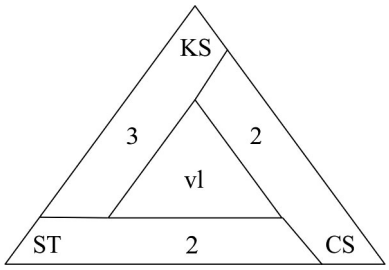

Рис. 7. Статистична модель визначення авторорозрізняльної здатності групи велярних фонем (KS – критерій Колмогорова-Смірнова, CS – критерій хі-квадрат, ST – критерій Стьюдента; vl – група велярних фонем; 2, 2, 3 – кількість порівнянь, в яких група велярних фонем має авторорозрізняльну здатність)

Отже, отримані результати тестування показали, що метод комплексного аналізу диференціації фоностатистичних структур авторського стилю дає змогу мінімізувати кількість груп фонем (група велярних фонем), за якими здійснюється авторська атрибуція тексту, і цим самим спростити процес авторської диференціації текстів.

Висновки:

1. Розроблено метод комплексного аналізу диференціації фоностатистичних структур авторського стилю. Метод грунтується на поєднанні критеріїв методу гіпотез (критеріїв Стьюдента, Колмогорова-Смірнова, хі-квадрат). Розроблений метод дає змогу підвищити достовірність диференціації стилів та змінімізувати кількість груп фонем (група велярних фонем), за якими здійснюється авторська атрибуція тексту.
2. Побудована статистична модель визначення авторорозрізняльної здатності груп приголосних фонем дає змогу визначити авторорозрізняльну здатність груп фонем і таким способом спростити процес авторської атрибуції тексту.
3. Розроблено програмну систему, яка дає змогу з більшою достовірністю продиференціювати тексти різних авторів. Отримані результати дають змогу ствердити, що група велярних фонем має високу авторорозрізняльну здатність у зіставленні текстів публічних промов американського президента Д. Трампа і статей С. Логан та Д. Вебстер з газети "Вільна газета".

Перелік використаних джерел

Altman, H. (2005). Moda ta istyna v linhvistytsi. *Problema kvantytatyvnoi linhvistyky*. Chernivtsi: Ruta, (pp. 3–11). [In Ukrainian].

Altmann, G., Kohler, R., & Piotrowski, R. (2005). *Quantitative Linquistik. Ein internationals Handbuch*. Berlin, New-York: de Gruyter.

Argamon, Sh., Koppel, M., Pennebaker, J., & Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM, 52*(2), 119–123. USA.

Bisikalo, O. V., & Vysotska, V. A. (2016). Sentence syntactic analysis application to keywords identification ukrainian texts. *Radio electronics computer science control, 3*(38), 54–65. Zaporizhzhya.

Gries, Th. S. (2009). *Statistics for Linguistics with R*. (Mouton Textbook), 335 p.

Juala, P. (2008). Authorship Attribution, Foundations and Trends (R) in *Information Retrieval, 1*(3), 233–334. Boston – Delft.

Khomytska, I. Yu., Tesliuk, V. M., & Labinska, L. S. (2018). Prohramna systema avtorskoi atrybutsii tekstiv na fonolohichnomu rivni. *Problemy ta perspektyvy rozvytku ekonomiky i pidpryiemnytstva ta kompiuternykh tekhnolohii v Ukraini: Collection of theses of the 14th scientific-practical conference*, Lviv. (pp. 15–16). [In Ukrainian].

Khomytska, I., Teslyuk, V., & Shakhovska, N. (Ed.). (2016). The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. *Advances in Intelligent Systems and Computing, 512*, 149–163. Lviv.

Khomytska, I., Teslyuk, V., & Shakhovska, N. (Ed.). (2018). Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level. *Advances in Intelligent Systems and Computing III, 871*, 105–118. Lviv.

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. Kharkiv.

Koppel, M. (2009). Computational Methods in Authorship Attribution. *Journal of the Association for Information Science and Technology, 60*(1), 9–26. USA.

Levitckii, V. V. (2007). *Kvantitativnye metody v lingvistike*. Vinnytsia: New Book, 259 p. [In Russian].

Lytvyn, V., Vysotska, V., Dosyn, D., Holoschuk, R., & Rybchak, Z. (2017a). Application of Sentence Parsing for Determining Keywords in Ukrainian Texts. *CSIT: Proceedings of the 12th Scientific and Technical Conference, Lviv*. (pp. 326–331).

Lytvyn, V., Vysotska, V., Pukach, P., Bobyk, I., & Uhryn, D. (2017b). Development of a method for the recognition of authors style in the ukrainian language texts based on linguometry, stylemetry and glottochronology. *Eastern-European Journal of Enterprise Technologies, 4/2*(88), 10–18.

Perebyinis, V. S. (2013). *Statystychni metody dlia linhvistiv*. Vinnytsia: New Book, 170 p. [In Ukrainian].

Piotrovskii, R. G. (1999). *Lingvisticheskii avtomat i ego rechemyslitelnoe obosnovanie*. Minsk, 126 p. [In Russian].

Segal, D. M. (1968). Statisticheskaia odnorodnost teksta na fonologicheskom urovne v polskom iazyke. *Strukturnaia tipologiia iazykov*, 85–93. Moscow. . [In Russian].

Segal, D. M. (1972). *Osnovy fonologicheskoi statistiki*. Moscow: Science, 255 p. [In Russian].

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

[1] Lviv Polytechnic National University, Lviv, Ukraine

[2] Ivano-Frankivsk National Technical University of Oil and Gas, Ivano-Frankivsk, Ukraine

MATHEMATICAL METHOD AND MODEL OF DIFFERENTIATION OF PHONOSTATISTICAL STRUCTURES OF AUTHORIAL STYLE

The method of complex analysis of differentiation of phonostatistical structures of authorial style of the English language has been developed. The method is based on a combination of the three statistical criteria: the Student's t-test, the Kolmogorov-Smirnov test and the chi-square test. The combination of the given criteria enables improving test validity of authorial styles differentiation. To solve the authorial styles differentiation problem, the statistical model, which allows improving test validity of authorship attribution of a text, has been built. The statistical model of author-differentiating capability for the velar phoneme group minimizes the number of phoneme groups by which the styles are differentiated. The developed program realizes the method and model using the Java programming language. This secures platform independence. To test the program, the speeches by D. Trump and the newspaper article by S. Lagon have been analyzed. The essential differences have been established in the stop and velar phoneme groups by the Student's t-test. The essential differences have been established in all eight phoneme groups by the Kolmogorov-Smirnov's test. The essential differences have been established in all phoneme groups except for stop and nasal phoneme groups by the chi-square t-test. Consequently, the greatest author-differentiating capability has been established in the velar phoneme group by the three tests applied. In another comparison, the papers by S. Logan and D. Webster from "Freedom Paper" have been differentiated. The essential differences have been established in the nasal, dorsal and velar phoneme groups by the Student's t-test. The essential differences have been established in all eight phoneme groups by the Kolmogorov-Smirnov's test. The essential differences have been established in sonorous, fricative, dorsal, stop, labial and nasal phoneme groups by the chi-square t-test. The combination of the three tests applied made it possible to establish the greatest author-differentiating capability in the dorsal phoneme group. Texts by different authors can be differentiated by this group. This minimizes the number of phoneme groups by which the authorship attribution is done.

Keywords: phonostatistical structure of a style; text differentiation; authorship attribution of a text.