Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.15421/40290627
Article received 28.05.2019 p.
Article accepted 27.06.2019 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

МЕТОД КОМПЛЕКСНОГО АНАЛІЗУ ДИФЕРЕНЦІАЦІЇ ФОНОСТАТИСТИЧНИХ СТРУКТУР СТИЛІВ АНГЛІЙСЬКОЇ МОВИ

Розроблено метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови. Метод ґрунтується на поєднанні двох статистичних критеріїв перевірки гіпотези на однорідність вибірки: критерію Стьюдента і критерію Колмогорова-Смірнова. Поєднання даних критеріїв забезпечує підвищення ефективності диференціації стилів. На основі розробленого методу, побудовано статистичну модель визначення ступеня дії чинника авторської манери викладу. Модель дає змогу підвищити ефективність стильової та авторської атрибуцій тексту. Розроблені метод і модель програмно реалізовано мовою програмування Java. POST запити двох типів: /process і /process/transcription. розроблено у програмі. За відсутності транскрипційного варіанта досліджуваного тексту, використовується перший запит, другий – за його наявності. Скоротити час роботи програми дає змогу другий запит. Вбудовану базу даних H2 написано мовою програмування Java. База даних H2 є відкритою, кросплатформною. Вона підтримує мову SQL, має добру інтеграцію із використовуваним фреймворком Spring Boot і не потребує додаткових інсталяцій. У структурі даних HashMap зберігається відповідь із сайту транскрипційного перекладу. Структура побудована на принципі ключ-значення і дає змогу уникати дублікатів. Якщо опрацьовується велика кількість даних, то зменшується кількість запитів у мережі Інтернет, що забезпечує незалежність і автономність програми. Малі затрати часу є характерними для роботи програми.

Ключові слова: фоностатистична структура стилю; стильова атрибуція тексту; авторська атрибуція тексту.

Вступ. За результатами аналізу методів, які використовуються для диференціації стилів та встановлення авторства досліджуваного тексту, з'ясовано, що методи математичної статистики є ефективними, бо мова є імовірнісною системою і її потрібно досліджувати імовірнісними методами. Задача встановлення авторства тексту передбачає диференціацію текстів на різних рівнях мови. Лексичний рівень є відкритою системою, в якому кількість елементів не є сталою. Система поповнюється новими словами, а рідко вживані слова виходять із системи. Авторський стиль відображає змінні процеси лексичної системи, які можна відобразити у моделюванні. Моделювання семантичних структур використано при текстовій диференціації (Shestakevych et al., 2014; Davydov, & Lozynska, 2016). Тексти продиференційовано за ключовими словами на лексичному та синтаксичному рівнях (Lytvyn et al., 2017a, 2017b). Однак ключові слова не дають вичерпної інформації про специфіку авторського стилю. Для розмежування текстів різних авторів, ефективним є відбір коротких слів (прийменник, сполучник), які характеризуються відносною стабільністю вживання (Bisikalo, & Vysotska, 2016). Через відкритість лексичної та синтаксичної систем, результати диференціації текстів на цих рівнях мови мають більш імовірнісний характер, ніж на фонологічному рівні (Burrows, 2002; Kapociute-Dzikiene et al., 2015; Stamatatos, 2009). Фонологічний рівень має сталу кількість елементів і є закритою системою. Тому точність розмежування текстів на фонологічному рівні є вищою, ніж на інших рівнях мови (Burrows, 2002; Kapociute-Dzikiene et al., 2015). Велику роль у підвищенні ефективності стильової та авторської диференціації текстів відіграє використання сучасних інформаційних технологій (ІТ). Однак, якщо інформаційні технології не застосовано на фонологічному рівні, то це позбавляє дослідження належного рівня точності (Argamon, 2009). Необхідно визначити поєднання ефективних квантитативних методів для диференціації текстів (Khomytska, & Teslyuk, 2016; Gries, 2009; Khomytska et al., 2018). Інструментальні засоби повинні реалізувати поєднання статистичних методів для забезпечення ефективності стильової та авторської атрибуції тексту (Koppel, 2009; Khomytska, & Teslyuk, 2018; Juala, 2008).

Аналіз поданих вище робіт показав, що актуальною є задача підвищення точності стильової і авторської атрибуції тексту та мінімізації кількості груп фонем (1–

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. Метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови. Науковий вісник НЛТУ України. 2019, т. 29, № 6. С. 140–143.
Citation APA: Khomytska, I. Yu., Teslyuk, V. M., & Beregovskyi, V. V. (2019). The method of complex analysis of differentiation of phonostatistical structures of english styles. *Scientific Bulletin of UNFU, 29*(6), 140–143. https://doi.org/10.15421/40290627

3 групи фонем), за якими розрізняються тексти. Метою дослідження є розроблення методики диференціації фоностатистичних структур стилів (Burrows, 2002; Stamatatos, 2009).

На основі зазначених наукових здобутків вперше розроблено метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови та багатофакторний метод визначення ступеня дії чинників стилю, підстилю та авторської манери викладу.

Методи та моделі диференціації фоностатистичних структур стилів. У дослідженні розроблено метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови, який ґрунтується на поєднанні методів гіпотез, ранжування і визначення відстаней між стилями (Khomytska et al., 2018a). Метод гіпотез передбачає застосування критеріїв Стьюдента і Колмогорова-Смірнова.

У плані алгоритму методу комплексного аналізу диференціації фоностатистичних структур стилів, наведемо основні математичні співвідношення. Вибірки перевірено на відповідність закону нормального розподілу за критерієм Пірсона: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$, $n_i' = \frac{N\Delta x}{S}\phi(Z)$, де n_i – кількість частот, які потрапляють в i-й інтервал.

Вибірки 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-ої і 2-ої вибірок за фіксованою групою фонем; n_1 і n_2 – кількість порцій 1-ої і 2-ої вибірок.

Вибірки продиференційовано за критерієм Колмогорова-Смірнова:

$$\lambda_{n,m} = \sqrt{\frac{nm}{n+m}} D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{-\infty<z<\infty} \left|F_n(z) - F_m(z)\right|, \qquad (1)$$

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.

Вибірки продиференційовано за методом ранжування: $r_{\bar{x}_1^a - \bar{x}_2^a} = r_{\max \bar{x}_1^a} - r_{\min \bar{x}_2^a}$ – різниця рангових показників 1-ої і 2-ої вибірок. Вибірки продиференційовано за методом визначення відстаней між стилями: $L = (t - t_0)/t$. Параметр t критерію Стьюдента відповідає величині типу $\bar{x}_1 - \bar{x}_0$, t_0 відповідно до вибраного рівня значущості.

Особливості реалізації програмної системи та результати дослідження. Розробленим програмним продуктом є програмна система (Khomytska et al., 2018b), структуру якої зображено на рис. 1.

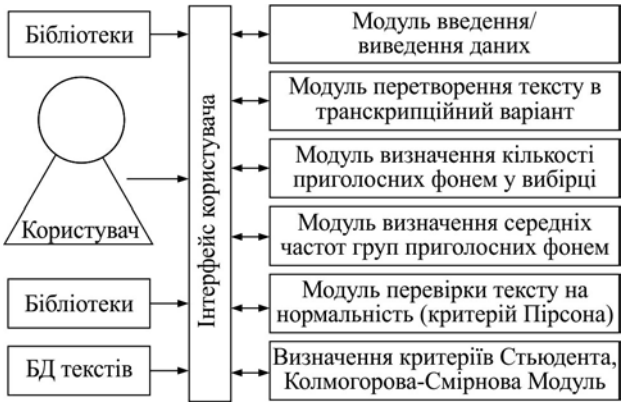

Рис. 1. Структура програмної системи диференціації фоностатистичних структур функціональних стилів англійської мови

Алгоритм функціонування системи диференціації фоностатистичних структур стилів передбачає:

Крок 1. Відбір текстів із чотирьох стилів і трьох підстилів та завантаження у файлах з розширенням *.txt.

Крок 2. Створення транскрипційних варіантів для всіх семи текстів стилів та підстилів.

Крок 3. Формування вибірки заданого обсягу за критерієм Стьюдента (31 тис. приголосних фонем).

Крок 4. Визначення частоти приголосних фонем у семи вибірках.

Крок 5. Поділ вибірки кожного із семи стилів та підстилів на 31 порцію.

Крок 6. Визначення частот та середнього значення частоти приголосних фонем по порціях для художнього, газетного, розмовного і наукового стилів.

Крок 7. Об'єднання фонем у 8 груп за акустико-артикуляційними ознаками.

Крок 8. Визначення середнього значення групи приголосних фонем у кожному стилі та підстилі.

Крок 9. Визначення теоретичного нормального розподілу на основі емпіричного нормального розподілу.

Крок 10. Визначення теоретичної частоти.

Крок 11. Перевірка відповідності частот груп приголосних фонем закону нормального розподілу за критерієм згоди Пірсона (1).

Крок 12. Диференціація двох зіставлених попарно вибірок за фіксованою групою приголосних фонем у семи текстах за критеріями Стьюдента (2) і Колмогорова-Смірнова (3).

Програма та алгоритм диференціації текстів реалізується у таких класах (рис. 3).

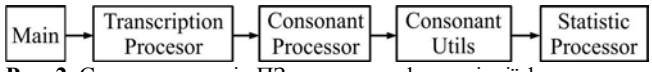

Рис. 2. Структура класів ПЗ системи диференціації фоностатистичних структур стилів

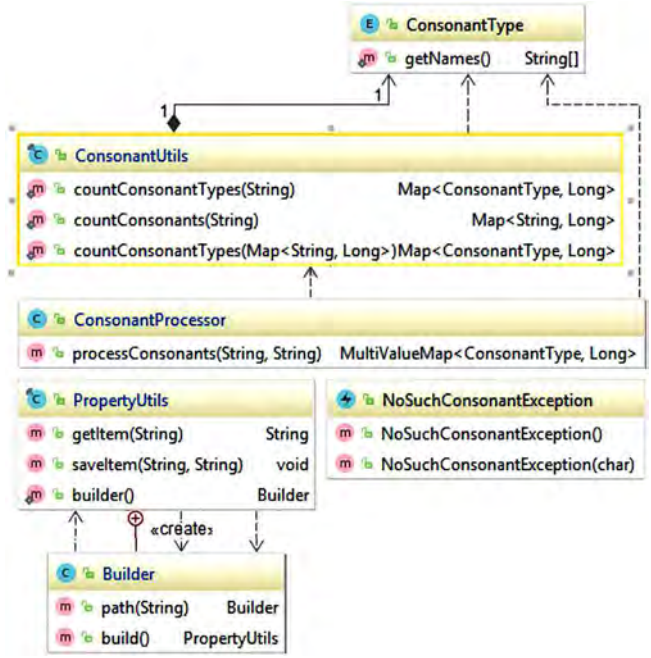

Рис. 3. Структура класів формування вибірки з приголосних фонем

Клас Main містить пакет flow, який забезпечує завантаження файлів, послідовність виконання операцій, отримання результатів оброблення. HTTP протокол використано для порівняння двох текстів. Через цей протокол здійснено завантаження двох файлів і вибір критерію перевірки гіпотези на однорідність вибірки. Клас

TranscriptionProcessor відповідає за перетворення тексту в його транскрипційний варіант. Текст розбивається на масив слів. Для кожного слова перевіряється його наявність у базі даних. Якщо слово відсутнє, за допомогою класу від java.net.HttpURLConnection, унаслідуваним від java.net.URLConnection, який підтримує з'єднання по мережевому протоколу HTTP, посилається запит на сайт транскрипційного перекладу http://upodn.com/phon.php.

Для формування вибірки з приголосних фонем створено клас ConsonantProcessor. На рис. 3 зображено UML діаграму, яка описує структуру класів формування вибірки з приголосних фонем.

Розроблення інформаційного забезпечення системи диференціації фоностатистичних структур функціональних стилів англійської мови. Вбудовану базу даних H2 написано мовою програмування Java. База даних H2 є відкритою, кросплатформною. Вона підтримує мову SQL, має добру інтеграцію із використовуваним фреймворком Spring Boot і не потребує додаткових інсталяцій.

У структурі даних HashMap зберігається відповідь із сайту транскрипційного перекладу. Структура побудована на принципі ключ-значення і дає змогу уникати дублікатів. Опрацьовується велика кількість даних, зменшується кількість запитів у мережі Інтернет, що забезпечує незалежність і автономність програми. Малі затрати часу є характерними для роботи програми. Для тестування програми вибрано тексти з публіцистичного стилю – статті С. Логан і Д. Вебстер з газети "Вільна Газета" ("Freedom Paper", papers by S. Logan and D. Webster). Результати, отримані за критерієм Стьюдента, зображено на рис. 4. Істотні відмінності встановлено за всіма вісьмома групами фонем. За результатами зіставлення текстів публіцистичного стилю – публічних промов американського президента Д. Трампа, статей С. Логан і Д. Вебстер з газети "Вільна Газета" побудовано статистичну модель визначення авторорозрізняльної здатності групи щілинних фонем (рис. 6).


```
    },
    "Statistic Results": {
        "SONOROUS": null,
        "FRICATIVE": null,
        "CORONAL": null,
        "DORSAL": false,
        "STOP": null,
        "LABIAL": null,
        "NASAL": false,
        "VELAR": false
    },
    "Error message": []
}
```


Рис. 4. Результати за критерієм Стьюдента


```
    },
    "Statistic Results": {
        "SONOROUS": null,
        "FRICATIVE": null,
        "CORONAL": null,
        "DORSAL": false,
        "STOP": null,
        "LABIAL": null,
        "NASAL": false,
        "VELAR": false
    },
    "Error message": []
}
```


Рис. 5. Результати за критерієм Колмогорова-Смірнова

Отже, результати тестування за методом комплексного аналізу диференціації фоностатистичних структур стилів англійської мови показали, що цей метод дає змогу змінімізувати кількість груп фонем (група щілинних фонем), за якими розрізняються тексти.

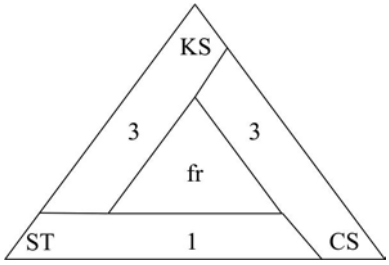

Рис. 6. Статистична модель визначення авторорозрізняльної здатності групи щілинних фонем (KS – критерій Колмогорова-Смірнова, CS – критерій хі-квадрат, ST – критерій Стьюдента; fr –група щілинних фонем; 1, 3, 3 – кількість порівнянь, в яких група щілинних фонем має авторорозрізняльну здатність.)

Висновки:

1. Розроблено метод комплексного аналізу диференціації фоностатистичних структур стилів англійської мови, який ґрунтується на поєднанні методів гіпотез (критерій Стьюдента, критерій Колмогорова-Смірнова), ранжування і визначення відстаней між стилями. Метод дає змогу підвищити точність диференціації стилів та змінімізувати кількість груп фонем (група щілинних фонем), за якими розрізняються тексти.
2. Побудовано статистичну модель визначення авторорозрізняльної здатності груп фонем, яка дає змогу визначити групу фонем з найвищою авторорозрізняльною здатністю, що спрощує процес авторської атрибуції тексту.
3. Розроблено програмні засоби, які дають змогу ефективніше здійснити стильову та авторську атрибуції тексту. Наведено результати, які дають змогу ствердити, що група щілинних фонем має найвищу авторорозрізняльну здатність у зіставленні текстів публіцистичного стилю – публічних промов американського президента Д. Трампа, статей С. Логан і Д. Вебстер з газети "Вільна Газета".

Перелік використаних джерел

Argamon, Sh. (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.

Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

Davydov, M., & Lozynska, O. (2016). Linguistic Models of Assistive Computer Technologies for Cognition and Communication. *Proceedings of the 11th Scientific and Technical Conference.* (pp. 171–174). Lviv.

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. *Boston–Delft, 1*(3), 233–334.

Kapociute-Dzikiene, J., Utka, F., & Sarkute, L. (2015). Authorship Attribution and Author Profiling of Lithuanian Literary Texts. *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, (pp. 96–105). Hissac, Bulgaria.

Khomytska, I. Yu., Tesliuk, V. M., & Labinska, L. S. (2018b). Prohramna systema avtorskoi atrybutsii tekstiv na fonolohichnomu rivni. *Problemy ta perspektyvy rozvytku ekonomiky i pidpryiemnytstva ta kompiuternykh tekhnolohii v Ukraini: zb. tez 14th nauk.-prakt. konf.*, (pp. 15–16). Lviv. [In Ukrainian].

Khomytska, I., & Teslyuk, V. (2016). The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspa-

per Styles on the Phonological Level. In *N. Shakhovska* (*Ed.*), *Advances in Intelligent Systems and Computing, 512*, 149–163. Lviv.
Khomytska, I., & Teslyuk, V. (2018). Authorship Attribution by Differentiation of Phonostatistical Structures of Styles. *CSIT: Proceedings of the 13th Scientific and Technical Conference*, (pp. 5–8). Lviv.
Khomytska, I., Teslyuk, V., & Shakhovska, N. (Ed.). (2018a). 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.
Lytvyn, V., Vysotska, V., Dosyn, D., Holoschuk, R., & Rybchak, Z. (2017b). Application of Sentence Parsing for Determining Keywords in Ukrainian Texts. *CSIT: Proceedings of the 12th Scientific and Technical Conference*. (pp. 326–331). Lviv.
Lytvyn, V., Vysotska, V., Pukach, P., Bobyk, I., & Uhryn, D. (2017a). 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.
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 9th Int. Conference CSIT2014*, (pp. 19–22), November 18–22, 2014. Lviv.
Stamatatos, E. (2009). A Survey of Modern attribution Methods. *Journal of the Association for Information Science and Technology, 60*, 538–556.

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

THE METHOD OF COMPLEX ANALYSIS OF DIFFERENTIATION OF PHONOSTATISTICAL STRUCTURES OF ENGLISH STYLES

The method of complex analysis of differentiation of phonostatistical structures of English styles has been developed. The method is based on a combination of the two statistical criteria of hypothesis verification for sample homogeneity: the Student's t-test and the Kolmogorov-Smirnov test. The combination of the criteria secures higher precision of style differentiation. On the basis of the developed method, the statistical model of determining the degree of author's style factor effect has been built. The model enables improving accuracy of style and authorship attribution. The developed method and model have been coded on the Java programming language. In the program, the POST requests of two types such as /process i/process/transcription have been developed. The first request is used when the transcription variant of the researched text is not available, the second – when it is available. The second request makes it possible to reduce the program operating time. An open, cross platform, inbuilt data base H2, written on the programming language Java has been used. The data base H2 supports the SQL and is well integrated with the used framework Spring Boot and doesn't need any additional installations. The reply from the transcription transformation site is stored in the data structure HashMap, built on the principle key-meaning and allows avoiding copies. The greater amount of data is processed, the fewer requests to the Internet are made. This secures program independence and autonomy. The program consumes little time. For testing the program, the texts from the publicist style have been chosen ("Freedom Paper", papers by S. Logan and D. Webster). 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 statistical model of author-differentiating capability for the fricative phoneme group has been built on the basis of the results obtained by the Student's t-test and by the Kolmogorov-Smirnov's test. The results of the program testing have shown that the method of complex analysis of differentiation of phonostatistical structures of English styles allows minimizing the number of phoneme groups by which the styles are differentiated.

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