

УДК 001.82:82

Юлія Калимон, Ігор Кульчицький, Ігор Ліхнякевич

ЕНТРОПІЯ БІГРАМ ЯК МЕТОД ФОРМАЛЬНОГО АНАЛІЗУ ХУДОЖНІХ ТВОРІВ

У статті простежено етапи становлення, окреслено основні об'єкти та проблеми застосування методів квантитативної лінгвістики у вивченні ідіолектів письменників на матеріалі їх художніх творів. Підкреслено значущість та актуальність цієї тематики у зв'язку із розвитком комп'ютерних технологій та програмного забезпечення, а саме: створення відповідного дослідницького корпусу. Зазначено, що згаданий метод дослідження художніх творів доповнюватиме літературознавчі студії та дасть змогу внести більше об'єктивності у традиційну методику. Новизна роботи полягає в застосуванні методу аналізу ентропії біграм як одного із перспективних у вивченні інформаційно-статистичних властивостей художніх текстів із метою якіснішого аналізу особливостей творчої манери письменників та порівняння із загальномовними показниками. Вперше таку методику дослідження застосовано для аналізу новел В. Стефаника та М. Яцківа, а також поезії П. Карманського. Запропоновано напрямки подальших досліджень на основі отриманих результатів.

Ключові слова: статистичні методи, частотність, біграма, ентропія, ідіолект, масив текстів.

Постановка наукової проблеми та її значення. Лінгвістика має різноманітні методи та засоби аналізу тексту. Загальноновизнаним є і той факт, що мова – структурно організована система знаків, що піддаються кількісному аналізу, результати якого дають змогу обґрунтувати чи відхилити висунуті гіпотези. Кількісні дослідження художніх творів не одразу увійшли до переліку традиційних методів досліджень як у лінгвістиці, так і в літературознавстві. Основними причинами недоречності таких методів для вивчення художніх творів називали нездатність вирішити питань суто літературознавчого характеру, ігнорування фонових явищ або акцент на технологічних аспектах, а також притаманна літературознавцям критичність до будь-якого підходу, який передбачає «об'єктивну» методологію [6; 9; 20; 23; 24; 26, с. 259–260]. Неабияку роль у покращенні ситуації відіграють можливості корпусів текстів та різноманітного програмного забезпечення. Статистичні методи можна застосовувати окремо для розв'язання певної мовознавчої проблеми, так і в комплексі з іншими якісними підходами. До сфер мовознавчого аспекту можна зарахувати як окремі дисципліни – математичну або статистичну лінгвістику, – так і деякі інші напрямки, де кількісні методики виступають як засоби розв'язання тих чи тих питань, пов'язаних із функціями мови: психолінгвістику, комп'ютерну лінгвістику, судову лінгвістику (лінгвістичну експертизу) та ін. Особливо актуальними такі дослідження стають у вивченні ідіолектів письменників, а також у питаннях атрибуції анонімних та псевдонімних текстів, визначення часу написання твору, для вивчення таких питань трактування тексту, як сюжет, тема, жанр, період, модальність тощо [2, 22; 22, с. 111].

Аналіз досліджень цієї проблеми. Загалом вивчення ідіолекту зводять до двох провідних проблем – особливості мови автора та атрибуція текстів [3, с. 22–23; 5, с. 181–183; 11, с. 338; 18, с. 378]. Поняття ідіолекту в лінгвістиці відоме давно та було предметом численних розвідок [6; 7]. Дослідниця Н. Журавльова, посилаючись на праці Г. Хердана та В. Вінтера, дає ще й таке узагальнене визначення авторського стилю: «це система періодично повторюваних вибірок, що характеризує індивідуальний спосіб вираження в мові конкретної людини» [4, с. 151–153]. Саме таке трактування передбачає застосування статистичних методів для аналізу авторського стилю. До проблеми атрибуції текстів (або стилеметрії) звертались ще до появи комп'ютерного інструментарію наприкінці XIX ст., проте із застосуванням математичних алгоритмів. Передусім це були дослідження, які стосувалися авторства античних творів, деяких творів Шекспіра тощо. Згодом, із 30-х років XX ст., стали широко застосовуватися і статистичні методи, однак авторів дослідження швидше цікавила сама можливість їх використання для визначення авторства чи особливостей стилю, аніж результат таких експериментів. З 70-х р. XX ст. все частіше математично-статистичні методи доповнюють можливостями комп'ютерної організації та опрацювання великих текстових масивів. Найпершим дослідженням такого плану став аналіз

роману «Тихий Дон» М. Шолохова. Подальший розвиток такої методики поширився і на проблему визначення авторства допитів та зізнань у випадках підозри на тиск на обвинуваченого з боку слідчих, що отримало широке розповсюдження в Америці та відоме кількома резонансними дослідженнями, що дало поштовх для розвитку такого напрямку, як лінгвістична експертиза [2, с. 23; 9, с. 140–143; 11, с. 118; 20; 22, с. 112; 27].

До питання вивчення ідіолекту письменника в докорпусний період зазвичай підходили суб'єктивно (т. зв. експертний метод) – філологи виокремлювали улюблені слова, терміни чи фразеологізми. Проте далеко не всі письменники послуговуються якоюсь особливою лексикою, яка б свідчила про особливості стилю. Відтак із появою корпусів текстів та все ширшим застосуванням статистичних методів (т. зв. формальні методи) розпочались пошуки альтернативних об'єктів дослідження. Аналіз літературних джерел дає змогу виокремити такі об'єкти: кількість складів у слові, довжина слів, довжина речень, розподіл повнозначних частин мови/службових частин мови, багатство словникового запасу (тут виокремлюють такі параметри, як коефіцієнт лексичної різноманітності, індекс Сімпсона (D), параметр Юла (K), ентропія та ін.), частотність слів, аналіз синтаксису (дерево залежностей) тощо. Водночас вчені зазначають, що жоден із них повністю не задовільняє критерій достовірності результатів із різних причин (величина досліджуваного тексту, недотримання принципу однорідності жанру та хронології творів тощо), а також наголошують на важливості використання кількох параметрів (напр., «золотий перетин» Г. Мартиненка, теорія розпізнавання образів М. Марусенка, кластерний, дискримінантний та факторний аналізи) [2, с. 24–25; 4, с. 151–153; 11, с. 125; 16, с. 137; 18, с. 380; 21, с. 339].

Ще одним перспективним методом називають статистичний аналіз, зокрема частотний, біграм та їх ентропії. Уперше його застосував та описав у своїй статті 1913 року російський вчений А. Марков на прикладі вивчення розподілу голосних та приголосних серед перших 20000 букв «Євгенія Онегіна», який згодом отримав назву ланцюг Маркова. На його основі Д. Хмельов зробив спробу аналізу творчого доробку російських письменників, де експериментально довів дієвість та ефективність цієї методики під час аналізу великої кількості авторів [17]. Цей метод є основою розроблених вченим програм розпізнавання та атрибуції текстів «Лінгвоаналізатор» та «Атрибутор». Ще одним прикладом застосування цього методу є праця групи вчених з Канади, які послуговувались досвідом В. Бенета 1976 року, що застосував цей метод на масиві зразків текстів Е. Хемінгвея, В. Шекспіра, Е. А. По та ін. Серед переваг такого підходу називають його незалежність від мови, яку аналізують, відсутність потреби у складному алгоритмі попереднього опрацювання тексту (напр., роззначення, частиномовна атрибуція) тощо [25, с. 256–257].

Російські вчені В. Гоголева та А. Шкарапута вводять у своє дослідження ще й поняття ентропії як «інтегральної характеристики тексту, яка може враховувати внесок кожного символу на основі частоти його вживання» [3, с. 23]. В інформатиці явище ентропії описує впорядкованість та складність матерії. За цією логікою текст є системою, а його ентропія досягає максимуму коли його символи трапляються з однаковою частотою. Згідно з працями А. Колмогорова та К. Шеннона ентропію тексту розглядають як «кількість інформації», яку несуть символи, без врахування змісту тексту [3, с. 23–24]. Така властивість корисна для визначення авторства твору в разі невеликого обсягу досліджуваного тексту або ж відсутності традиційних об'єктів дослідження мови автора (використання фразеологізмів, улюблених слів тощо). На основі таких міркувань було здійснено обрахунки ентропії та частотності символів низки російських письменників із подальшою інтерпретацією результатів для визначення дати написання творів та їх атрибуції [3, с. 24–27].

Мета і завдання статті. Завдання нашого дослідження – застосування методу аналізу частотності біграм (тут – послідовності букв) та їх ентропії як один із прикладів застосування формальних математичних методів вивчення особливостей стилю письменника. Метою є проаналізувати результати обчислювань на матеріалі новел В. Стефаника [12; 13; 14; 15] та М. Яцківа [19], а також поезії П. Карманського [8] й дійти висновків щодо ефективності цього методу.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Об'єктом дослідження обрано твори саме цих авторів як представників одного хронологічного періоду – кінець XIX – початок XX ст. - та жанру (новели В. Стефаника та М. Яцківа). Поезія П. Карманського використана для порівняльної характеристики прозового й поетичного текстів. Спочатку тексти дослідницького корпусу перетворено в електронний формат. Після цього їх

нормалізовано, поділено на абзаци і речення. У текстах залишено тільки символи розширеної української абетки, до якої входять символи традиційного алфавіту, пробіл, дефіс та апостроф; усі букви перетворено у великі; між словами залишено лише один пробіл; утворено один суцільний текст послідовним приєднанням через пробіл. У такий спосіб здійснено обрахунок загальної кількості рядків та символів. Загальну характеристику творів подано в таблиці 1.

Таблиця 1

Загальна кількість рядків та символів у масивах текстів

Автор	Кількість рядків	Кількість символів
Василь Стефаник	2649	286092
Михайло Яцків	2286	246888
Петро Карманський	2025	218700

Подальший алгоритм дій окреслено нижче:

Крок 1. У кожному масиві вибрано послідовно десять сегментів із інтервалом в 0,02 довжини тексту. Для кожного сегмента обраховано кількість рядків, біграм та їхню ентропію, визначено їх середнє значення.

Крок 3. У кожному масиві вибрано десять випадкових сегментів, які корелюють зіставні за розміром із послідовними сегментами. Для кожного сегмента обраховано кількість рядків, біграм та їхню ентропію та визначено їх середнє значення.

Далі подано таблиці з узагальненими результатами обрахунків для творів В. Стефаника (Табл. 2), М. Яцківа (Табл. 3) та П. Карманського (Табл. 4)

Таблиця 2

Результати дослідження ентропії біграм новел В. Стефаника

В. Стефаник	Кількість		Ентропія
	рядків	біграм	
Весь текст	2649	288740	111,48
Послідовний сегмент (середнє)	212	23108	104,00
Випадковий сегмент (середнє)	264	28301	106,09

Таблиця 3

Результати дослідження ентропії біграм новел М. Яцківа

М. Яцків	Кількість		Ентропія
	рядків	біграм	
Весь текст	2286	249173	111,89
Послідовний сегмент (середнє)	183	19947	106,13
Випадковий сегмент (середнє)	228	24440	107,22

Таблиця 4

Результати дослідження ентропії біграм поезії П. Карманського

П. Карманський	Кількість		Ентропія
	рядків	біграм	
Весь текст	2025	220724	109,91
Послідовний сегмент (середнє)	162	17658	105,75
Випадковий сегмент (середнє)	202	21655	106,03

На Рис. 1 подано зведені обрахунки ентропії біграм у досліджуваних сегментах масивів текстів. Зображені сегменти 1-10 – це послідовні сегменти, а 11–20 – випадково вибрані.

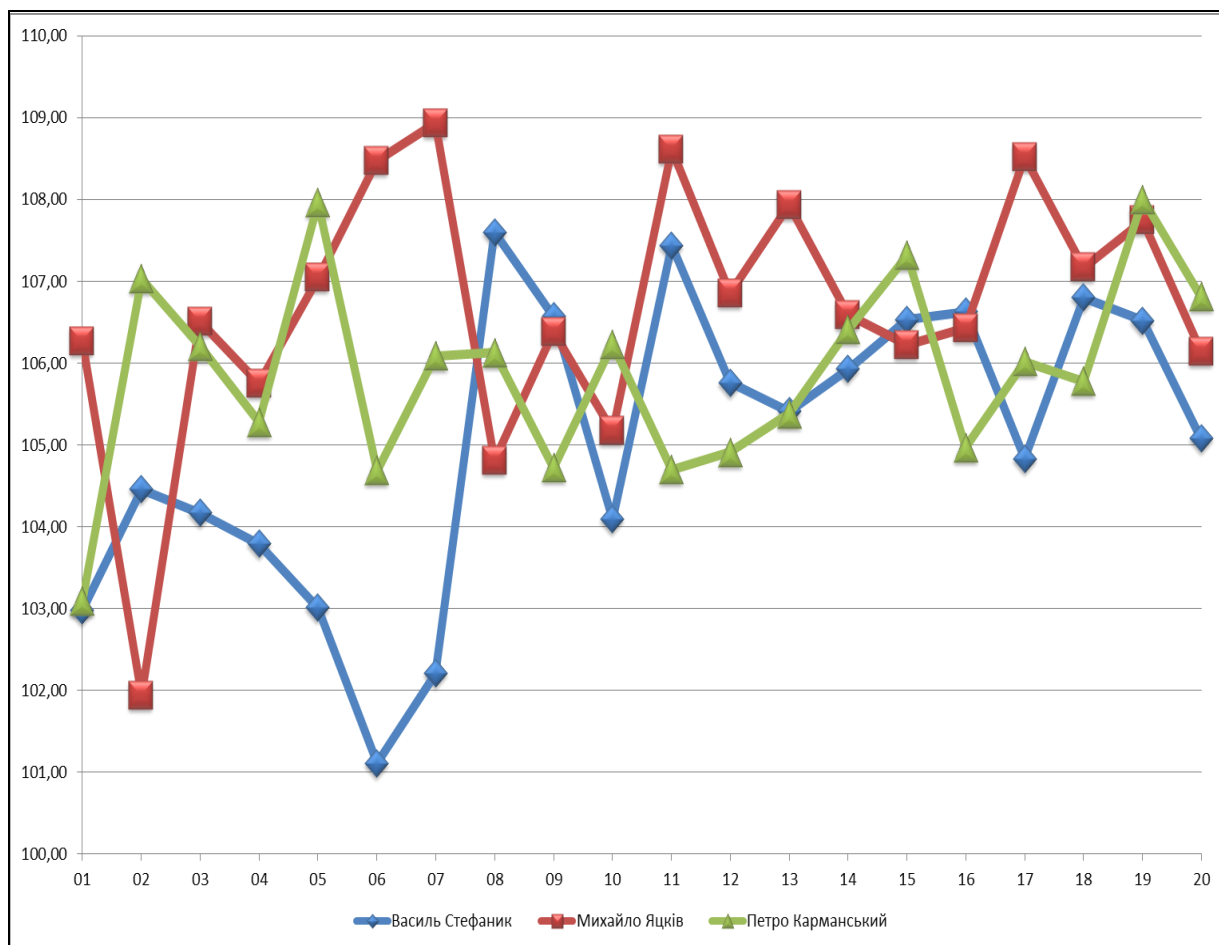


Рис. 1. Значення ентропії біграм у масивах текстів В. Стефаника, М. Яцків та П. Карманського

Аналіз отриманих даних показує, що ентропія біграм в усіх масивах текстів має приблизно однакове значення (із незначною відмінністю у В. Стефаника), незважаючи на наявність поетичних творів поряд із прозовими. Окрім цього, значення ентропії для випадково вибраних сегментів достатньо стабільне.

Висновки та перспективи подальшого дослідження. Отримані результати на матеріалі творів В. Стефаника, М. Яцків та П. Карманського дають змогу дійти таких висновків: 1) ентропія біграм не може бути авторським інваріантом; 2) для статистичних досліджень краще робити вибірки у випадковий, а не послідовний спосіб. Щодо подальших досліджень, то вважаємо доцільним продовжити подібні обрахунки на рівні триграм і т. ін., а також на матеріалі творів інших письменників задля перспективи отримання узагальнених висновків щодо всієї української мови.

Джерела та література

1. Батура Т. В. Формальные методы установления авторства текстов и их реализация в программных продуктах / Т. Батура // Программные продукты и системы. – Изд-во : ЗАО НИИ «Центрпрограммсистем». – Тверь, 2013. – Вып. 4. – С. 286–295.
2. Верховин С. С. К вопросу о лингвотеретических основах методик авторизации текста / С. С. Верховин // Учёные записки Забайкальского госуд. ун-та. Серия : Филология, история, востоковедение. – № 2 (49). – 2013. – С. 22–27.

3. Гоголева В. А. Математический подход к установлению авторства и времени создания текста на основе исследования его энтропии / В. А. Гоголева, А. П. Шкарапута // Вестник Пермского университета. Серия : Математика. Механика. Информатика. – Пермь, 2014. – № 4. – С. 22–28.
4. Журавлёва Н. Н. Применение количественных методов при анализе стиля автора и решении проблем атрибуции / Н. Н. Журавлёва // Вестн. Тюменского гос. ун-та. – 2012. – № 1. – С. 150–155.
5. Захаров В. Н. Программная система поддержки атрибуции текстов статей Ф. М. Достоевского / В. Н. Захаров, А. А. Леонтьев, А. А. Рогов, Ю. В. Сидоров // Труды Петрозаводского госуниверситета. Сер. : Прикладная математика и информатика. – Петрозаводск : ПетрГУ, 2000. – Вып. 9. – С. 180–189.
6. Калимон Ю. О. Співвідношення термінів на позначення системи мовних засобів письменника / Ю. О. Калимон // Актуальні проблеми філології та перекладознавства. Вип. 12. – Хмельницький НУ, 2017. – С. 92–96.
7. Калимон Ю. О. Ідіолект, ідіостиль, індивідуальний стиль. Тотожне чи різне? / Ю. О. Калимон, І. М. Кульчицький, І. О. Ліхнякевич // Науковий вісник ВНУ ім. Лесі Українки. Серія : Філологічні науки. Мовознавство. – Луцьк, 2014. – № 6. – С. 226–229.
8. Карманський П. Поезії / П. С. Карманський. – К. : Укр. письменник, 1992. – 361 с.
9. Кульчицький І. М. Дослідження довжини речення та слова у творах Романа Іванчука / Ігор Кульчицький // Вісник Нац. ун-ту «Львівська політехніка». – Львів, 2017. – С. 139–148.
10. Кульчицький І. М. Технічні аспекти функціонування текстів у електронному інформаційному просторі / Ігор Кульчицький // Український інформаційний простір. Число 2. – Київський національний університет культури і мистецтв. – Київ 2014. – С. 101–108.
11. Родионова Е. С. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика : межвуз. сб. / под ред. А. С. Герда. – СПб : Из-во С.-Пб. ун-та, 2008. – Вып. 7. – С. 118–127.
12. Стефаник В. Межа / Василь Стефаник // Літературно-науковий вісник. – Т. 92, кн. 2. – Львів, 1927. – С. 97–98.
13. Стефаник В. Портрет : твори / Василь Стефаник. – Львів : ДВУ, 1929. – С. 94–95.
14. Стефаник В. Твори [з дереворитами В. Касіяна і М. Бутовича] / Василь Стефаник. – Львів : з друкарні Видавничої Спілки «Діло», 1933. – 222 с.
15. Стефаник В. Шкільник / Василь Стефаник // Рідна школа. – Львів, 1932. – № 1. – С. 2–4.
16. Суровцева Т. Г. О построении статистических критериев для атрибуции авторства литературных текстов / Т. Г. Суровцева, С. П. Чистяков. – Вестник СПбГУ, Сер. 10, Вып. 3. – 2009. – С. 137–142.
17. Хмельов Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д. В. Хмельов // Вестник МГУ. Сер. 9: Филология. – № 2. – 2000. – С. 115–126.
18. Хозяинов С. А. Некоторые проблемы и методы квантитативно-структурного изучения авторских стилей / С. А. Хозяинов // Известия Рос. Госуд. педагог. ун-та им. А. И. Герцена. – 2008. – № 63(1). – С. 378–383.
19. Яцків М. Вибрані твори / Михайло Яцків. – Київ : Дніпро, 1973. – 453 с.
20. Grieve J. W. Quantitative Authorship Attribution: A history and an evaluation of techniques / J. W. Grieve. – Simon Fraser University. – 2005. – 282 p.
21. Holmes D. I. The Analysis of Literary Style-A Review / D. I. Holmes // Journal of the Royal Statistical Society. Series A (General), vol. 148, no. 4. – 1985. – P. 328–341.
22. Holmes D. I. 'The Evolution of Stylometry in Humanities scholarship / D. I. Holmes // Literary and Linguistic Computing Vol.13(3). – 1998. – P. 111–117.
23. Hoover D. Quantitative Analysis and Literary Studies. In: A Companion to Digital Literary Studies / D. Hoover. [Electronic resource]. — Access mode : http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.id=0&brand=9781405148641_brand
24. Jones S. When Computers Read: Literary Analysis And Digital Technology / Jones S. // Bulletin of the American Society for Information Science and Technology. - Vol. 38. - No 4. – 2012. - P. 27–30.
25. Kešelj V. N-gram based author profiles for authorship attribution / V. Kešelj, F. Peng, N. Cercone, C. Thomas // Proceedings of the conference Pacific Association for Computational Linguistics, PACLING. – Vol. 3. – 2008. – P. 255–264.
26. Siemens R. A New Computer-assisted Literary Criticism? / R. Siemens // Computers and the Humanities. – 2002. – P. 259–267.
27. Wright D. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem / D. Wright // International Journal of Corpus Linguistics. – Vol. 22(2). – 2017. – P. 212–241.

References

1. Batura, T. V. 2013. "Formalne metody ustanovleniia avtorstva tekstov i ikh realizatsiia v programmnykh produktakh". *Prohrammnye produkty i sistemy*, 4: 286–295.
2. Verkhozin, S. S. 2013. "K voprosu o lingvotoreticheskikh osnovakh metodik avtorizatsii teksta". *Uchenye zapiski Zabaikalskogo gosudarstvennogo universiteta. Seriya: Filologiya, istoriya, vostokovedenie*, 2: 22–27.
3. Gogoleva, V. A., and Shkaraputa, A. P. 2014. "Matematicheskii podkhod k ustanovleniyu avtorstva i vremeni sozdaniia teksta na osnove issledovaniia ego entropii". *Vestnik Permskogo universiteta. Seriya: Matematika. Mekhanika. Informatika*, 4: 22–28.
4. Zhuravliova, N. N. 2012. "Primenenie kolichestvennykh metodov pri analize stilia avtora i reshenii problem atributsyi". *Vestnik Tiimenskogo gosudarstvennogo universiteta*, 1: 150–155.

5. Zakharov, V. N., and Leontiev, A. A., and Rohov, A. A., Sidorov, Yu. V. 2000. "Prohrammnaya sistema podderzhki atributsyi tekstov statey F. M. Dostoyevskogo". *Trudy Petrozavodskogo gosudarstvennogo universiteta*, 9: 180–189.
6. Kalymon, Yu. O. 2017. "Spivvidnoshennia terminiv na poznachennia systemy movnykh zasobiv pysmennyka". *Aktualni problemy filolohii ta perekladoznavstva*, 12: 92–96.
7. Kalymon, Yu. O., and Kulchytskyi, I. M., and Likhniakievych, I. O. 2014. "Idiolekt, idiostyl, indyvidualnyi styl. Totozhne chy rizne?". *Naukovy visnyk VNU im. Lesi Ukrayinky. Seriya :Filolohichni nauky. Movoznavstvo*, 6: 226–229.
8. Karmanskyi, Petro. 1992. *Poeziyi, vstupne slovo V. I. Luchuka*. Kyiv: Ukr. pysmennyk.
9. Kulchytskyi, I. M. 2017. "Doslidzhennia dovzhyny rechennia ta slova u tvorax Romana Ivanychuka". *Visnyk Natsionalnoho universytetu "Lvivska politekhnikha"*. Seriya: Informatsiini systemy ta merezhi, 139–148.
10. Kulchytskyi, I. M. 2014. "Tekhnichni aspekty funktsionuvannia tekstiv u elektronnomu informatsiinomu prostori". *Ukrayinskyi informatsiinyi prostir*, 2: 101–108.
11. Rodionova, E. S. 2008. "Metody atributsyi khudozhestvennykh tekstov". *Strukturnaya i prikladnaya lingvistyka*, 7: 118–127.
12. Stefanyk, Vasyl. 1927. "Mezha". *Literaturno-naukovy visnyk*, 92: 97–98.
13. Stefanyk, Vasyl. 1929. "Portre". *Tvory, peredmova V. Koriaka*, 94–95. Lviv : DVU.
14. Stefanyk, Vasyl. 1933. *Tvory, z derevorytamy V. Kasiiana i M. Butovycha*. (Lviv: z drukami Vydavnychoi Spilky «Dilo».
15. Stefanyk, Vasyl. 1932. "Shkilnyk". *Ridna shkola*, 1: 2–4.
16. Surovtseva, T. H., and Chystyakov, S. P. 2009. "O postroyenii statisticheskikh kriteriev dlia atributsii avtorstva literaturnykh tekstov". *Vestnik SPbGosU*, 10:137–142.
17. Khmeliov, D. V. 2000. "Raspoznavaniye avtora teksta s ispolzovaniyem tsepej A. A. Markova". *Vestnik MGU, Filologiya*, 2: 115–126.
18. Khoziainov, S. A. 2008. "Nekotorye problemy i metody kvantitativno-strukturnogo izucheniia avtorskikh stylei". *Izvestiya Rossiiskogo gosudarstvennogo pedagogicheskogo universiteta im. A. Y. Gertsena*, 63: 378–383.
19. Yatskiv, Mykhailo. 1973. *Vybrani tvory*. Kyiv: Dnipro.
20. Grieve, J. W. 2005. *Quantitative Authorship Attribution: A history and an evaluation of techniques*. Simon Fraser University.
21. Holmes, D. I. 1985. "The Analysis of Literary Style A Review". *Journal of the Royal Statistical Society. Series A (General)*, 148: 328–341.
22. Holmes, D. I. 1998 "The Evolution of Stylometry in Humanities scholarship". *Literary and Linguistic Computing* 13(3): 111–117.
23. Hoover, D. 2008. *Quantitative Analysis and Literary Studies*. In: *A Companion to Digital Literary Studies*, edited by Susan Schreiban and Ray Siemens. Oxford: Blackwell. http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.id=0&brand=9781405148641_brand
24. Jones, S. 2012. "When Computers Read: Literary Analysis And Digital Technology". *Bulletin of the American Society for Information Science and Technology*, 38: 27–30.
25. Kešelj, V., and Peng, F., and Cercone, and Thomas, N. C. 2008. "N-gram based author profiles for authorship attribution". *Proceedings of the conference Pacific Association for Computational Linguistics, PACLING*, 3: 255–264.
26. Siemens, R. 2002. "A New Computer-assisted Literary Criticism?" *Computers and the Humanities*, 259–267.
27. Wright, D. 2017. "Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem". *International Journal of Corpus Linguistics*, 22: 212–241.

Калымон Юлия, Кульчицкий Игорь, Лихнякевич Игорь. Энтропия биграмм як метод формального аналізу художественних текстів. В статтю прослідковані етапи становлення, обозначены основні об'єкти і проблеми в застосуванні методів квантитативної лінгвістики в вивченні ідіолекта письменників на матеріалі їх творів. Серед найбільш розповсюджених методів називають вирахування кількості слогів в слові, довжини слів, довжини речень, розподіл знаменальних частин мови / службових частин мови, багатство словникового запасу і др. В той же час все актуальніше стають і такі об'єкти дослідження, як биграми і їх ентропія. Інтерес до них обумовлюють можливістю аналізу незначительних текстових фрагментів, а також відсутністю необхідності в затратному і трудомісткому процесі попередньої підготовки текстів (розметка, частеречової атрибуції). Підкреслено значимість і актуальність цієї тематики в зв'язі з розвитком комп'ютерних технологій і програмного забезпечення, а саме створення відповідного дослідницького корпусу. Зазначено, що згаданий метод зможе доповнювати дослідження художественних творів і дозволить внести більше об'єктивності в традиційну методіку. Новизна роботи заключається в застосуванні методу аналізу ентропії биграмм як одного з перспективних в вивченні інформаційно-статистичних властивостей художественних текстів з метою якісного аналізу особливостей творчої манери письменників і її порівняння з загальноязыковими показателями. Вперше така методика дослідження застосована для аналізу творчості новелл В. Стефаніка і М. Яцківа, а також поезії П. Карманського. Такий вибір обумовлений однорідністю хронологічного (для всіх трьох авторів) і жанрового критеріїв (для новелістів В. Стефаніка і М. Яцківа). Поезія П. Карманського вибрана для проведення паралелів між прозаїчними і поетичними творами. Представлено напрями подальших досліджень на основі отриманих результатів.

Ключові слова: статистичні методи, частотність, биграма, ентропія, ідіолект, масив текстів.

Kalymon Yuliia, Kulchytskyi Ihor, Likhniakievych Ihor. The Entropy of Bigrams as Formal Analysis Method of Literary Texts. The article describes stages of formation, main objects and problems in the application of methods of quantitative linguistics in the study of the idiolect of writers on the material of their works. According to the literature analysis among the most common methods are the calculations of the number of syllables in a word, word length, sentence length, distribution of parts of speech / function words, the richness of vocabulary, etc. At the same time such objects of research like bigrams and their entropy are becoming more relevant nowadays. It is caused by the possibility of analyzing minor text fragments, as well as by the absence of the need for a costly and time-consuming process of preliminary preparation of texts (markup, part-of-speech attribution). The significance and relevance of this subject was stressed in connection with the development of computer technologies and software, namely, the creation of an appropriate research corps. It is noted that this method of research of literature will be able to complement literary studies and will allow bringing more objectivity into traditional methods. The novelty of the work is the application of the method of entropy analysis of bigrams as one of the promising in the study of information and statistical properties of literary texts with the purpose of qualitative analysis of the features of the creative manner of writers and its comparison with general language indicators. For the first time such technique was used to analyze novellas by V. Stefanyk and M. Yatskiv, as well as the poetry of P. Karmanskyi. This choice of authors is due to the homogeneity of the chronological (for all three authors) and genre criteria (for the novelists V. Stefanyk and M. Yatskiv). Poetry by P. Karmanskyi was chosen to draw parallels between prose and poetic works. Further research perspectives based on the results are outlined.

Key words: statistical methods, frequency, bi-gram, entropy, idiolect, array of texts.

УДК 811.111'373.613+821.111(71)'06-3.09Грекул

Ірина Калиновська

УКРАЇНІЗМИ В АНГЛОМОВНІЙ КАНАДСЬКІЙ ЛІТЕРАТУРІ (НА ПРИКЛАДІ РОМАНУ ЛІЗИ ГРЕКУЛ «ПІСНЯ КАЛИНИ»)

Статтю присвячено вивченню запозичень українського походження, висвітлено аспекти їхнього вживання в англomовній канадській літературі. Аналіз відображає чинники, які впливають на використання українізмів, їхні семантичні особливості, способи запозичення та асиміляції на прикладі роману канадської письменниці українського походження Лізи Грекул «Пісня Калини». Українізми відображають реалії, унікальні для української національної культури, історії, способу життя, а також функціонують як маркер української національної ідентичності, українськості, яку можна розуміти як один з аспектів канадійності в сучасній мультикультурній Канаді, адже вона пов'язана з поточними дослідженнями питань етнічності, національності та мультикультуралізму в канадській літературі. Українськість головних героїв роману підкреслюється авторкою як національними символами України, так і вживанням українізмів у творі. Виявлені українізми проаналізовано і розподілено на семантичні групи, зокрема топоніми, антропоніми та соціально-побутову лексику. До найбільш поширених українізмів належать слова на позначення українських національних страв, музичних інструментів, а також антропоніми. Розглянуто їхню фонетичну й граматичну асиміляцію. У перспективі увага повинна бути зосереджена на вивченні факторів, які впливають на особливості вживання українізмів у творах інших англomовних письменників українського походження, різних варіантах англійської мови, їхні особливості у порівняльному аспекті.

Ключові слова: запозичення, українізм, українськість, семантична група, спосіб запозичення, асиміляція.

Постановка наукової проблеми та її значення. Мова є важливою для національної ідентичності, але необхідно зазначити, що процеси мовної та етнічної ідентифікації українців в Канаді різняться. Так, Ю. Жлуктенко у своїй монографії «Українська мова на лінгвістичній карті Канади» зазначає, що коли в Канаді не залишиться жодної людини, яка б розмовляла українською мовою, можна запевняти, що ніякої асиміляції власне не сталося: мова та культура зникли, зате залишилася «ідентичність» [3]. Тобто, навіть майже повністю втративши свою мову, канадці українського походження продовжуватимуть ідентифікувати себе українцями.

Жити за межами України – означає адаптуватися та асимілюватися в новому суспільстві, а це веде до того, що носії української мови або забудуть з часом рідну мову повністю, або прилаштують її структуру та лексику до нових соціокультурних умов Канади, або використовуватимуть українську мову диференційовано до певних ситуацій та потреб. В