системы. – ІПММС НАНУ, 2000. – N 1. – С. 120–127. 9. Кульчицький І. М. Технічні аспекти опрацювання комп'ютером природномовної інформації / І. М. Кульчицький // Вісник Нац. ун-ту "Львівська політехніка": Інформаційні системи та мережі. – 2014. – № 783. – С. 344–353. 10. Родионова Е. С. Методы атрибуции художественных текстов / Е. С. Родионова [Електронний ресурс]. – Режим доступу: http://epir.ru/pragmat!/projects/corneille/files/ Metody_atributsii.pdf 11. Ружевич Н. А. Математична статистика: навч. посібник для студ. базового напряму "Прикладна математика" / Н. А. Ружевич. – Л.: Видавництво Національного ун-ту Львівська політехніка, 2001. – 167 с. 12. Статистичні параметри стилів / за ред. В. С. Перебийніс. – К.: Наукова думка, 1967. – 260 с. 13. Фрумкина Р. М. Роль статистических методов в современных лингвистических исследованиях // С. К. Шаумян Математическая лингвистика / отв. ред. С. К. Шаумян. – М.: Наука, 1973. – С. 156–183. 14. Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д. В. Хмелёв. – М.: Вестник МГУ, 2000. – сер.9: филология, № 2. – С.115–126. 15. Черемшина Марко. Твори: у 3 т. / Повне видання за ред. Є. Пеленського. – Львів: Ізмарагд, 1937. – Т. 1: 206 с. – Т. 2.: 245 с. – Т. 3.: 231 с. 16. Широков В. А. Інформаційна теорія лексикографічних систем / В. А. Широков. – К.: Довіра, 1998. – 331 с. 17. Ihor Kulchytskyy Idiolect of Marko Cheremshyna: quantitative aspect / Kulchytskyy Ihor, Sozanska Khrystyna // Комп'ютерні науки та інформаційні технології: Матеріали IX Міжн. конф. CSIT 2014. – Львів: Видавництво Львівської політехніки, 2014. – С. 143–145.

УДК 811.161.2'33:519.25

І. М. Кульчицький, У. С. Шандрук
Національний університет "Львівська політехніка",
кафедра прикладної лінгвістики

ВПЛИВ ОРФОГРАФІЇ НА ЧАСТОТНІСТЬ БУКВ У ТЕКСТАХ

Розглянуто один із важливих напрямків квантитативних досліджень мови та мовлення – вивчення інформаційно-статистичних властивостей тексту. Здійснено спробу перевірки на творах Леся Мартовича впливу орфографії на відносну частотність букв у текстах. Зроблено відповідні висновки.

Ключові слова: квантитативні дослідження, частота букв, орфографія, відносна частота, Лесь Мартович.

The article is dedicated to one of the most important areas of quantitative studies of language and speech that is the study of information and statistical properties of text. On the basis of works by Les Martovych an attempt was made to verify impact of spelling on the relative frequency of letters in the text. A number of relevant conclusions were made.

Key words: quantitative study, frequency of letters, spelling, relative frequency, Les Martovych.

Вступ

Одна з проблем науки – вибір адекватних методів дослідження, адже останні, поряд з об'єктом та предметом, становлять обов'язкову складову пізнання у будь-якій галузі людської діяльності. Стосуються ці проблеми і лінгвістики, де, починаючи від середини минулого століття, поширені математичні, зокрема квантитативні, методи вивчення системи мови та особливостей її функціонування [4, 29]. Квантитативні методи поділяють на кількісні та статистичні, зводячи кількісні до простих підрахунків вжитку тих чи інших одиниць мови, а статистичні – до

використання апарату теорії ймовірності та математичної статистики. На думку авторів, такий поділ навряд чи доцільний, позаяк будь-яке практичне використання теоретичних формул передбачає кількісні підрахунки. Один з сучасних напрямків застосування статистичних методів – стилістичні особливості як окремих творів, так і всієї творчості автора, що зумовлює актуальність цього дослідження.

Загальна постановка проблеми

У сучасному мовознавстві існує ідея системності мовних фактів [6, 18]. Проте, якщо визнання системного характеру фонетики та граматики не викликає сумнівів, то щодо системності лексики ще донедавна були суперечки. Під сумнів ставлять сам факт її системності [6, 18]. Супрун зазначає: "… для успішного функціонування мови як засобу комунікації її лексика мусить бути організована. Іншими словами, лексика не може не бути системною, якщо словниковий запас мови успішно використовують в процесі її функціонування."

Відомо, що об'єктивні властивості мови виявляються як в якісних, так і в кількісних ознаках. Без використання кількісних характеристик та статистичних методів не можна пояснити функціонування низки мовних категорії [18, 94]. Згідно з Будуеном-де-Куртене: "Потрібно частіше використовувати в мовознавстві кількісне, математичне мислення, і таким чином наближаючи його все ближче до точних наук" [1]. Тоді як фонетика, соціолінгвістика, експериментальна психолінгвістика користуються методами точних наук вже протягом тривалого часу, такі як дисципліни як морфологія, синтаксис, семантика набули статистичного спрямування лише від 1990-х років XX ст. Стрімкий розвиток статистичних методів у лінгвістиці був зумовлений такими чинниками: обробка емпіричних даних передбачала більше поширення статистичних інструментів, почали з'являтися нові погляди на дослідження мовних явищ, які ґрунтувалися на теорії ймовірностей, теорії інформації, статистичному моделюванні, досягненнях комп'ютерної та корпусної лінгвістики. В результаті кількісні методи в лінгвістиці стали стійкою методологічною основою [34].

Одним із найпоширеніших методів дослідження лексики певного автора, твору чи жанру є статистичний, адже кількісні характеристики тексту дозволяють встановити не лише склад лексики, але й співвідношення використання її різних пластів, співвідношення слів, які зустрічаються рідко та часто і т. д. Статистичні методи дослідження дають змогу зробити об'єктом дослідження весь склад, так би мовити, нейтральної лексики, яка є показником різноманітності чи одноманітності словника письменника [18, 96].

Першими спробами застосування статистики в лінгвістичних дослідженнях були частотні словники, які подають списки слів з частотою їх використання в певному тексті. Першим таким словником вважають словник Ф. Кединга "Частотний словник німецької мови", 1898 р. [18, 97]

З іншого боку, у зв'язку зі стрімкою комп'ютеризацією та розвитком статистичного моделювання мови почав зростати інтерес до атрибуції текстів [12]. Проблему встановлення авторства анонімних та псевдонімних текстів активно вивчають та досліджують. Вважають, що кожен автор має стиль написання, властивий тільки йому [33, 150], й на підставі якісних та кількісних характеристик його стилю можна зробити висновок щодо авторства.

Перші успішні методи атрибуції з'явилися в кінці XIX століття, проте досі їх використовують лише епізодично [12], адже як наука атрибуція перебуває у стані становлення та розвитку. Методи атрибуції ще до кінця не досліджено [12]. Піонером у цій галузі можна вважати фізика-теоретика Томаса Менденхолла (Thomas Corwin Mendenhall), який вперше показав, що деякі прості статистичні методи можуть виявитися корисними для вирішення питань щодо спірного авторства [Mendenhall]. Він провів аналогію зі спектрометрією. Проаналізувавши певний твір, можна отримати гістограму частоти вживання слів різної довжини. Він вважав, що частота вживання слів кожного автора буде різною, а відтак гістограма частоти стане унікальною, що дасть можливість правильно встановити авторство. Отже, математична статистика допомагає перетворити суб'єктивний метод на об'єктивну техніку [33, 152].

Атрибуція тексту зумовлює низку труднощів. Насамперед стиль автора не залишається незмінним впродовж його життя, тому доводиться досліджувати часові залежності характеристик стилю. Стиль залежить від жанру: автор може імітувати діалекти персонажів, що також ускладнює атрибуцію. Статистичний аналіз передбачає попереднє опрацювання тексту: його потрібно розбити на однорідні за жанром частини, забрати власні назви і т. д. Для точнішої атрибуції потрібно враховувати додаткову інформацію про текст, кандидатів та час написання твору.

Ефективність атрибуції проявляється в двох випадках: спростування кандидата на авторство та поява нового кандидата, якого раніше не брали до уваги [30].

Незважаючи на величезну різноманітність описаних методів, жоден з них ніколи не застосовували до великої кількості текстів. Річ у тім, що часто ці методи не піддаються автоматизації і вимагають деякого людського втручання, що призводить до практичної неможливості обробки великої кількості текстів великого обсягу. Тому постає питання спільності кожного з методів: чи можна застосовувати будь-який з них поза ситуації, до якої їх було розроблено [10, 96]?

Новий метод визначення авторства текстів, написаних природною мовою, вперше запропонував Хмелев у роботі [30]. Він ґрунтується на формальній математичній моделі послідовності букв (і будь-яких інших елементів) тексту як реалізації ланцюга Маркова. За творами відомих авторів обчислюють матрицю перехідних частот вживання пар елементів (букв, граматичних класів слів і т. д.). Вона є оцінкою матриці ймовірності переходу з елемента в елемент. Матрицю перехідних частот будують для кожного учасника. Для кожного автора оцінюють ймовірність того, що саме він написав анонімний текст (або фрагмент тексту). Автором анонімного тексту вважають того, у кого обчислена оцінка ймовірності найбільша (тобто використовується принцип максимальної правдоподібності). Такий метод, як показує його перше використання [30] в додатку до різноманітного матеріалу, демонструє свою дивовижну точність [10, 97]. Тобто, використання такої, здавалося б, простої одиниці, як пара букв, які йдуть у тексті поспіль, дає точніші результати, ніж використання таких мовних категорій, як поодинокі граматичні класи слів та їхні пари. Підрахунок частот вживань пар букв надає інформацію про словник автора, а також подекуди інформацію про граматичні конструкції, яким він надає перевагу [10, 105]. Під час атрибуції текстів слід пам’ятати про те, що твори одного і того самого автора могли бути видані у різний час різними правописами. Мета цього дослідження – дослідити вплив зміни орфографії на частотність букв у одному і тому самому тексті. Як приклад розглянуто творчість Леся Мартовича.

Аналіз досліджень та публікацій

Застосування математичних методів у лінгвістиці передбачили у своїх творах ще Ф. де Соссюр та І. А. Бодуен де Куртене. Однак поширилися вони із виходом робіт з машинного перекладу лише з 50-х років XX сторіччя [5, 95]. Серед українських дослідників мови статистичними дослідженнями займаються такі науковці, як Володимир Широков [31], Максим Кригін [9], Валентина Перебийніс [27], Соломія Бук [3] та ін. Серед зарубіжних учених, що займались чи займаються цією тематикою [2; 3], згадаємо Петера Ґжибека (Peter Grzybek) та Еммеріха Келіха (Emmerich Kelih) з Австрії, Ґабріелю Альтман (Gabriel Altmann) та Рейнгарда Коелера (Reinhard Köhler) з Німеччини, Адама Павловскі (Adam Pawłowski) та Ядвігу Самбор (Jadwiga Sambor) з Польщі, Гейзу Віммер (Geiza Wimmer) з Словаччини, Юхана Тулдаву з Естонії, Раймунда Піотровского та Анатолія Шайкевича з Росії та ін.

Аналіз наукових результатів

Становлення та розвиток українського правопису

Перший український правописний узус ґрунтувався на Кирило-Мефодієвській традиції, пристосованій до вимог східнослов’янських мовних систем. У світській практиці він зберігався протягом XII–XVI ст. Так званий другий південнослов’янський орфографічний (графіко-орфографічний) вплив помітно позначився лише на правописі конфесійного письменства XVI ст. і мало торкнувся світського [6]. Характерно, що саме для запису світських текстів, насамперед

офіційно-ділового стилю, староукраїнської мови, виникла потреба шукати засоби для позначення фонем ґ, дж, дз, і такі засоби було знайдено.

Другий етап в історії правопису пов'язаний із "Граматикою" (1619 р.) М. Смотрицького. Незважаючи на те, що автор нормалізував правопис (та орфоепію) церковнослов'янської мови української редакції, орфографія вченого в підхожих пунктах доволі послідовно застосовувалася в українських текстах, зокрема у світській літературі: в ній вживалася буква ґ у загальних та власних назвах [20].

Новоукраїнські правописні системи (а їх було близько 50) [21] в силу відомих історичних причин не знаходили офіційного затвердження аж до кінця XIX ст. Серед них слід згадати правопис М. Смотрицького 1619 року, правопис за І. Котляревським, правопис О. Павловського 1818 року, правопис М. Максимовича 1827 року, правопис "Русалки Дністрової" 1837 року, кулішівка 1856 року, правопис Київський 1873 року.

У 1876 році правопис було заборонено, а його послідовний розвиток припинено силою. Згодом було зроблено спроби докорінно змінити правопис. Вважали, що одним із більших недоліків українського правопису було вживання окремих поодиноких значків для так званих йотованих голосних, що складалися з двох звуків, – замість писати ja, je, ji, ju пишемо я, є, ї, ю. Учений М. Драгоманів у своїх Женевських виданнях 1877 року писав уже по-новому: јама, моју, даје, стојшть јавыр над водоју. Таким чином з'явилася драгоманівка 1877 року, яка ненадовго поширилась у Галичині. Також драгоманівкою писав ідеологічний учень Драгоманова І. Франко [26].

Подальший розвиток правопису пов'язуємо із Галичиною, де австрійський уряд не пригнічував розвиток культури, а тому правопис міг розвиватися нормально. У 1886 році вийшов відомий "малорусско-німецкий словар" Євгена Желехівського, який значно вплинув на усталення та запровадження фонетичного правопису в Галичині. До 1905 року чинним був фонетичний правопис, яким писав Борис Грінченко – так звана грінченківка.

Етапною у справі впорядкування українського правопису вважаємо працю, здійснену під егідою Наукового товариства ім. Т. Шевченка, яке аж до 1918 р. відігравало роль Української академії наук. У НТШ 1900 р. створено "Язикову комісію" для вироблення норм українського правопису, до якої в різні роки входили О. Колесса, М. Павлик, В. Гнатюк, І. Кокорудз, К. Студинський, С. Смаль-Стоцький, І. Франко, М. Грушевський, В. Дорошенко, А. Кримський, Є. Тимченко [7]

Виняткову роль у впорядкуванні українського правопису в XX ст. відіграла орфографічна система, що її застосував у "Словарі української мови" (1907–1909 рр.) Б. Грінченко [25].

У 1919 році було скликано Правописну комісію, в якій професор І. Огієнко подав на розгляд свої раніше складені "Правила українського правопису". Це була перша наукова система українського правопису. Комісія складалася із проф. А. Є. Кримського, проф. Є. К. Тимченка, проф. Н. Грунського, О. Курилова, Г. Голоскевича та інших.

17 січня 1919 р., міністр народної освіти проф. Іван Огієнко ухвалив складений правописний кодекс для обов'язкового вжитку в усій Україні, і він вийшов у світ під назвою "Головніші правила українського правопису". Це була основа майбутнього академічного правопису [26].

У 1920 році Всеукраїнська Академія наук знову переглянула ці "Правила" й ухвалила їх до загального вжитку з деякими доповненнями. Так з'явився перший авторитетний правописний кодекс в Україні – академічний правопис.

У 1926–1927 рр. найгостріше, як і нині, стояла проблема вживання букв *ґ, ґ* та твердого й м'якого *л* в іншомовних словах, тобто в ділянці, де найбільше відрізнялися західно- та східноукраїнська орфографічні традиції [8].

З'являлися сумніви, що правопис – справа не самої Академії наук, а всього громадянства, і що правопис слід переглянути. Народний комісар освіти Микола Скрипник прихилився до цієї думки, й у травні 1927 р. у Харкові відбулася конференція. У Харківській Правописній комісії найбільше сперечань було за вимову чужоземних слів, особливо за західноукраїнську, що різко відрізнялася від східноукраїнської вимови.

Всеукраїнська академія наук (ВУАН) використала принципи Б. Грінченка та І. Огієнка при опрацюванні “Найголовніших правил українського правопису”. Ці правила стали основою для опрацювання першого загальноукраїнського правопису, що вийшов у Харкові 1928 року і був запроваджений у практику з 1 січня 1929 року.

“Український правопис” 1928 р. – перша спроба сформувати єдиний, соборний орфографічний кодекс для народу. Обговорення його правил відбулося на всіх територіях проживання українців як в органах масової інформації, так і на сторінках наукових журналів. Його прийнято демократичним шляхом – голосуванням на всеукраїнській конференції, на якій були присутні представники різних українських земель [20].

У 1933 році виходить новий правопис, з якого було вилучено майже все, що відбивало оригінальні риси української фонетико-морфологічної системи. Цей правопис був запроваджений без жодних обговорень, директивно. Написання багатьох слів відповідало російським формам. Надалі таке “очищення” тривало. На початку 1942 року радянський уряд в Україні доручив Академії наук поновити свою роботу з упорядкування українського правопису. В жовтні 1942 року Академія наук схвалила складений правопис і передала його на затвердження радянському урядові. Уряд не поспішав і доручив перегляд складеного правопису ще й новій Комісії, в яку ввійшли академіки: Л. А. Булаховський, П. Г. Тичина й М. Т. Рильський та письменник Ю. І. Яновський, головував у Комісії заступник Голови народних комісарів М. П. Бажан. Зрештою народній комісар освіти П. Г. Тичина прийняв новий правопис і вніс його на затвердження й Раді народних комісарів, яка остаточно затвердила його 8 травня 1945 року [26].

У редакції 1945 р. нічого з викинутих правил 1928 р. не відновлено. Деякі елементи далі наближено до російського правопису, зокрема введено флексію -і в родовому відмінку однини іменників із – ен-: “імені” замість “імени” [8]. В кінці 50-х років було створено об’єднану правописну комісію, а у листопаді чергову редакцію “Українського правопису” підписано до друку, і книжка вийшла в 1960 р. Українські правила наближено до “Правил русской орфографии й пунктуации”, виданих 1956 р. [18]. Ця редакція кодексу була чинною майже 30 років [26].

У часи так званої перебудови радянського суспільства виникла ідея внести деякі зміни в редакцію 1960 р. У 1988 р орфографічна комісія при Відділенні літератури, мови й мистецтвознавства АН УРСР приступила до чергового редагування й доповнення “Українського правопису”. 14 листопада 1989 р. було затверджено п’яту редакцію “Українського правопису”, опубліковану 1990 року. Найпомітнішою в цій редакції є повернення в алфавіт букви ґ (хоч і з дуже обмеженою сферою вжитку: у власне українських і давно засвоєних словах із інших мов) [28].

Українська еміграція за кордоном та діаспора ніколи не визнавали брутальних змін 1933 р. в нормах “Українського правопису” 1928 р., через це в орфографічній практиці двох частин нації стався розкол [8]. На хвилі національно-державницького руху почали лунати голоси про перегляд норм правопису, а на І Міжнародному конгресі україністів, що відбувався в Києві протягом 27 серпня — 3 вересня 1991 р., заявлено про потребу вироблення для всіх українців світу єдиних орфографічних норм. Ідею опрацювання таких норм, що мають ґрунтуватися на всьому історичному досвіді їх творення, підтримала Орфографічна комісія Академії наук [28].

У 1999 році під керівництвом В. В. Німчука було розроблено проект реформи українського правопису – скорочена назва – Проєкт. До нього внесено передусім ті зміни, що враховують столітні традиції української орфографії [24].

Вибір та організація матеріалу дослідження

Дослідницьку збірку текстів створено на основі першого тому повного зібрання творів Леся Мартовича [15] та трьох його збірок видання 1903 [17], 1904 [16] і 1922 [14] років. Матеріал було вибрано таким чином тому, що твори 1903–1905 років за правописом сильно відрізняються і від сучасного правопису, і від правопису 1922 року. Останній, своєю чергою, від сучасного теж значно відрізняється. Оскільки в українських текстах окрім букв у їхньому статусі використовують знаки дефіса, апострофа та пробілу (останній поділяє текст на слова), а основний предмет цього дослідження – відносна частота вживання букв української абетки, то під час обчислень тексти

творів Леся Мартовича інтерпретувались як множина символів розширеної української абетки, до якої, окрім її букв, додано знаки апострофа, дефіса та пробілу. Підготовка матеріалу мала два кроки.

Крок 1. Тексти всіх творів перетворено на електронну форму та нормалізовано [11]. Тексти творів було розбито на 4 групи. До першої та другої груп увійшли відповідно однойменні твори 1903–1905 років (перша) та 2011 року (друга): “Відміна”; “Війт”; “Гарбата”; “Грішниця”; “Зле діло”; “Іван Рило”; “Кадриль”; “Квіт на п’ятку”; “Лумера”; “Ось поси моє!”; “Прощальний вечір”; “Стрибожий дарунок”; “Хитрий Панько”.

До третьої та четвертої груп – однойменні твори 1922 (третя) року та 2011 (четверта) року видання: “Булка”; “Мужицька смерть”; “На торзі”; “Не читальник”; “Нічний гість”; “Перша сварка”; “Ян”.

Крок 2. Для проведення дослідження всі твори чотирьох груп було видозмінено за такими правилами [16]:

- усі малі букви українського алфавіту замінено великими;
- символи тексту, що не входять до розширеної абетки, замінено пробілом;
- між словами залишено лише один пробіл;
- під час обчислень символ кінця абзацу рахували як пробіл.

Організація та результати дослідження

Дослідження проводили на персональному комп’ютері класу IBM PC під управлінням операційної системи Windows у програмних середовищах Python та MS Office.

Спочатку для всіх чотирьох груп було пораховано абсолютні та відносні частоти символів розширеного алфавіту. Результати подано у табл. 1, 2.

Таблиця 1

Абсолютні частоти символів розширеного алфавіту у групах творів Леся Мартовича

Символ	Група творів			
	1	2	3	4
1	2	3	4	5
А	30474	30358	16310	16360
Б	7564	7574	3456	3464
В	17710	17744	8786	8748
Г	5318	5410	2918	2930
Ґ	230	198	178	170
Д	11510	11662	5718	5776
Е	16496	16236	8636	8554
Є	1926	1392	1202	1256
Ж	3580	3594	2284	2300
З	7072	7140	3604	3592
И	21808	21626	10892	11122
І	9986	14548	7492	7538
Ї	5396	1418	642	650
Й	5524	5862	3258	3206
К	11886	11890	6120	6146
Л	11366	11374	5572	5608

Продовження табл. 1

1	2	3	4	5
М	9410	9420	4808	4822
Н	16552	16778	8288	8326
О	30920	31446	15114	15074
П	9054	9050	4508	4532
Р	12972	13006	6604	6640
С	12954	12380	6454	6300
Т	16682	16848	9360	9364
У	10674	10740	5256	5294
Ф	96	96	94	78
Х	3448	3426	1762	1768
Ц	1648	2176	1330	1510
Ч	3904	3946	1922	1936
Ш	2818	2968	1586	1962
Щ	2092	1940	1242	878
Ь	4392	4360	2436	2426
Ю	2554	2546	1272	1248
Я	7444	7618	4096	4126
Апостроф	48	308	24	174
Дефіс	292	650	316	334
Пробіл	66636	67214	35054	35546
Разом	**382436**	**384942**	**198594**	**199758**

Таблиця 2

Відносні частоти символів розширеного алфавіту у групах творів Леся Мартовича

Символ	Група творів			
	1	2	3	4
1	2	3	4	5
А	0,079683921	0,078863829	0,082127355	0,081899098
Б	0,019778473	0,019675691	0,017402338	0,017340983
В	0,046308402	0,046095256	0,044241014	0,043792990
Г	0,013905595	0,014054065	0,014693294	0,014667748
Ґ	0,000601408	0,000514363	0,000896301	0,000851030
Д	0,030096539	0,030295473	0,028792411	0,028914987
Е	0,043134015	0,042177783	0,043485705	0,042821814
Є	0,005036137	0,003616129	0,006052549	0,006287608
Ж	0,009361043	0,009336471	0,011500851	0,011513932
З	0,018491983	0,018548249	0,018147577	0,017981758
И	0,057023920	0,056179892	0,054845564	0,055677370
І	0,026111559	0,037792706	0,037725208	0,037735660
Ї	0,014109550	0,003683672	0,003232726	0,003253937
Й	0,014444247	0,015228268	0,016405329	0,016049420
К	0,031079710	0,030887770	0,030816641	0,030767228
Л	0,029720005	0,029547308	0,028057242	0,028073970
М	0,024605424	0,024471219	0,024210198	0,024139208
Н	0,043280444	0,043585787	0,041733386	0,041680433
О	0,080850129	0,081690229	0,076105018	0,075461308

Продовження табл. 2

1	2	3	4	5
П	0,023674549	0,023510035	0,022699578	0,022687452
Р	0,033919401	0,033786908	0,033253774	0,033240221
С	0,033872334	0,032160689	0,032498464	0,031538161
Т	0,043620370	0,043767633	0,047131333	0,046876721
У	0,027910552	0,027900307	0,026466056	0,026502068
Ф	0,000251022	0,000249388	0,000473327	0,000390472
Х	0,009015888	0,008900042	0,008872373	0,008850709
Ц	0,004309218	0,005652800	0,006697080	0,007559147
Ч	0,010208244	0,010250895	0,009678037	0,009691727
Ш	0,007368553	0,007710252	0,007986143	0,009821884
Щ	0,005470196	0,005039720	0,006253965	0,004395318
Ь	0,011484274	0,011326382	0,012266232	0,012144695
Ю	0,006678242	0,006613983	0,006405027	0,006247560
Я	0,019464695	0,019789994	0,020624994	0,020654993
Апостроф	0,000125511	0,000800121	0,000120850	0,000871054
Дефіс	0,000763526	0,001688566	0,001591186	0,001672023
Пробіл	0,174240919	0,174608123	0,176510871	0,177945314

Теоретичним підгрунтям наступної частини дослідження був критерій згоди К. Пірсона (χ^2) [23]. За гіпотетичну теоретичну функцію розподілу почергово прийнято відносний частотний розподіл збірок творів першої та третьої груп. Як експериментальні дані прийнято відповідно абсолютну частоту символів у творах другої та четвертої груп. Як нульову гіпотезу H_0 прийнято твердження: “при зміні орфографії розподіл частот символів розширеної української абетки не змінюється”. Для її перевірки виконано такі кроки:

- попарно порівняно розподіл частоти символів першої з другою та третьої з четвертою;
- отримані частоти використано для обчислення статистики критерію $\chi^2_{експ.}$;
- $t_{кр} = \chi^2_{1-\alpha,k-1} = 49.802$ визначали за відповідною таблицею [23] за рівня значущості α=0,05 та ступенях свободи для k=36 (кількість символів розширеної української абетки);
- якщо отримували, що $\chi^2_{експ.} \geq t_{кр}$, то гіпотезу відхиляли, інакше її приймали.

Результати обчислених нормованих відхилень [23] та $\chi^2_{експ.}$ подано у табл. 3.

Таблиця 3

Результат обчислень під час перевірки гіпотези H_0

Символ	Групи			
	перша з другою		третя з четвертою	
	P(x)	Нормоване відхилення	P(x)	Нормоване відхилення
1	2	3	4	5
А	0,0796839	3,2490014	0,0821274	0,1267261
Б	0,0197785	0,2056038	0,0174023	0,0432124
В	0,0463084	0,3776504	0,0442410	0,9063235
Г	0,0139056	0,6102205	0,0146933	0,0088721
Ґ	0,0006014	4,8496460	0,0008963	0,4567676
Д	0,0300965	0,5061707	0,0287924	0,1042415
Е	0,0431340	8,1602210	0,0434857	2,0246505
Є	0,0050361	154,1271364	0,0060525	1,8235516
Ж	0,0093610	0,0248285	0,0115009	0,0029720
З	0,0184920	0,0659025	0,0181476	0,3026613

Продовження табл. 3

1	2	3	4	5
И	0,0570239	4,8089720	0,0548456	2,5200319
І	0,0261116	2011,5625852	0,0377252	0,0005785
Ї	0,0141096	2965,5649903	0,0032327	0,0278013
Й	0,0144442	16,3815914	0,0164053	1,5424063
К	0,0310797	0,4562986	0,0308166	0,0158269
Л	0,0297200	0,3862923	0,0280572	0,0019920
М	0,0246054	0,2817752	0,0242102	0,0415806
Н	0,0432804	0,8292367	0,0417334	0,0134212
О	0,0808501	3,3602843	0,0761050	1,0876058
П	0,0236745	0,4400685	0,0226996	0,0012940
Р	0,0339194	0,1992192	0,0332538	0,0011035
С	0,0338723	33,2948905	0,0324985	5,6683667
Т	0,0436204	0,1913761	0,0471313	0,2747601
У	0,0279106	0,0014477	0,0264661	0,0097879
Ф	0,0002510	0,0040952	0,0004733	2,8972109
Х	0,0090159	0,5729879	0,0088724	0,0105662
Ц	0,0043092	161,2594379	0,0066971	22,1666360
Ч	0,0102082	0,0685964	0,0096780	0,0038686
Ш	0,0073686	6,0995980	0,0079861	84,2927770
Щ	0,0054702	13,0403838	0,0062540	110,3424336
Ь	0,0114843	0,8356351	0,0122662	0,2405507
Ю	0,0066782	0,2380072	0,0064050	0,7733333
Я	0,0194647	2,0927374	0,0206250	0,0087160
Апостроф	0,0001255	1395,7817928	0,0001208	930,2898444
Дефіс	0,0007635	431,4116669	0,0015912	0,8203606
Пробіл	0,1742409	0,2978928	0,1765109	2,3286218
$\chi^2_{експ.}$		7221,6382408		1171,1814544

Як бачимо, гіпотезу H_0 відхилено в обох випадках.

Висновки та перспективи наукових досліджень

У результаті проведених досліджень стосовно творів Леся Мартовича можна дійти таких висновків:

Заміна орфографії суттєво впливає на частоту символів у текстах.

Для української мови частота таких символів, як “І”, “И” та апостроф хоча і суттєво впливає на частотний розподіл, частота інших символів теж вносить свою вагому лепту.

Отримані результати стосуються лише творчості Леся Мартовича. Для узагальнень щодо інших українськомовних художніх творів необхідні подальші дослідження.

Дослідження проведено лише для частотного розподілу поодиноких символів. Експерименти необхідно продовжити для біграм, триграм і т. д.

1. Бодуэн де Куртенэ И. А. Избранные труды по общему языкознанию: В 2. т. / И. А. Бодуэн де Куртенэ. – М., 1963. 2. Бук С. Лінгвостатистичний опис “Не спитавши броду” Івана Франка / Соломія Бук [Електронний ресурс]. – Режим доступу: http://www.lnu.edu.ua/faculty/Philol/www/visnyk/55_2011/55_2011_Buk.pdf 3. Бук С. Сучасні методи дослідження мови письменника у слов’янознавстві / С. Бук [Електронний ресурс]. – Режим доступу: http://www.lnu.edu.ua/page/n61/010.pdf 4. Верхозин С. С. О статусе количественных методов в лингвистике / С. С. Верхозин [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/o-statuse-kolichestvennyh-metodov-v-lingvistike 5. Гладкий А. В. “Математические методы изучения естественных языков”, Математическая логика, теория алгоритмов и теория множеств, Сборник

работ. Посвящается академику Петру Сергеевичу Новикову к его семидесятилетию, Тр. МИАН СССР, 133, 1973, 95–108 [Електронний ресурс]. – Режим доступу: http://www.mathnet.ru/links/ 3ff1f6b395ed41df319615fb89072f40/ tm2737.pdf 6. Гнатенко Л. А. Староукраїнський правопис останньої чверті XIV – першої чверті XVII ст. у зв'язку з проблемою другого південнослов'янського графіко-орфографічного впливу (букви на позначення голосних звуків) / Л. А. Гнатенко. – К., 1997. 7. Гузар О. Правописна концепція НТШ і питання уніфікації єдиного українського правопису на зламі XIX–XX століть / О. Гузар // Український правопис та реалії сьогодення / Матеріали засідань Мовознавчої комісії та Комісії всесвітньої літератури НТШ у Львові 1994–1995 рр. – Львів, 1996. – 13 с. 8. Історія українського правопису: XVI–XX століття. Хрестоматія. – К.: Наукова думка, 2004. – 584 с. 9. Кригін М. Ю., Широков В. А. Дослідження інформаційно-статистичних властивостей українського тексту / М. Ю. Кригін, В. А. Широков // Математические машины и системы. – ІПММС НАНУ, 2000. – N 1. – С. 120–127. 10. Кукушкина О. В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелёв // Проблемы передачи информациию. – 2001. – Т. 37, Вып. 2. – С. 96—109. 11. Кульчицький І. М. Технічні аспекти опрацювання комп'ютером природномовної інформації / І. М. Кульчицький // Вісник Нац. ун-ту "Львівська політехніка": Інформаційні системи та мережі. – 2014. – № 783. – С. 344—353. 12. Малютов М. Б. Обзор методов и примеров атрибуции текстов / М. Б. Малютов. – Бостон: Северо-восточный университет США, 2005. – 40 с. 13. Марков А. А. Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цеп / А. А. Марков // Изв. Имп. акад. наук. Сер. 6. – 1913 г. – № 3. – С. 153–162. 14. Мартович Л. Не-читалник и иньші оповідання. Вид. 2. – Львів: 1922. – 96 с. 15. Мартович Л. Повне зібрання творів. Т. 1. / За редакцією Г. Марчук. – Івано-Франківськ, ЛІК, 2011. – Т.1. – 340 с. 16. Мартович Л. Стрибожий дарунок і иньші оповідання. – Львів: Будзиновський, 1905. – 96 с. 17. Мартович Л. Хитрий Панько і иньші оповідання. – Львів: Вид. Спілка, 1903. – 103 с. 18. Методы изучения лексики / под ред. А. Е. Супруна. – Минск: Издательство БГУ им. В. И. Ленина, 1975 г. – 232 с. 19. Москаленко А. А. Історія українського правопис (радянський період) / А. А. Москаленко. – Одеса, 1968. – С. 47. 20. Німчук В. В. Проблеми українського правопису в XX ст / В. В. Німчук // Літературна Україна. – 2001. 21. Огієнко І. Нариси з історії української мови: система українського правопису / І. Огієнко // Популярно-науковий курс з історичним освітленням. – Варшава, 1927. – С. 16–20. 22. Родионова Е. С. Методы атрибуции художественных текстов / Е. С. Родионова // Структурная и прикладная лингвистика. – С.-Петерб. ун-т, 2008. – Вып. 7. – С. 118–127. 23. Ружевич Н. А. Математична статистика: навч. посібник для студ. базового напряму "Прикладна математика" / Н. А. Ружевич. – Л.: Видавництво Нац. ун-ту Львівська політехніка, 2001. – 167с. 24. Русанівський В. М. Стосунок "проєкту" до реального українського правопису / В. М. Русанівський // Мовознавство. – 2002. – № 6. – С. 92–98. 25. Словарь украинского языка, собранный редакцией журнала "Киевская старина"/ Редактировал, с добавлением собственных материалов, Б. Д. Гринченко. – К., 1907. – С. 23. 26. Стан української літературної мови. XVIII. Історія українського правопису. – [Електронний ресурс]. – Режим доступу: http://litopys.org.ua/ohukr/ohu20.htm 27. Статистичні параметри стилів / за ред. В. С. Перебийніс. – К.: Наукова думка, 1967. – 260 с. 28. Український правопис. – 4-те вид., виправл. і доповн. – К., 1993. – С. 6. 29. Фрумкина Р. М. Роль статистических методов в современных лингвистических исследованиях // С. К. Шаумян. Математическая лингвистика / ответственный редактор С. К. Шаумян. – Москва: Наука, 1973. – С. 156–183. 30. Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д. В. Хмелёв // Вести МГУ. Сер. 9. Филология. – 2000. – С. 115–126. 31. Широков В. А. Інформаційна теорія лексикографічних систем / В. А. Широков. – К.: Довіра, 1998. – 331 с. 32. Cantos G. P. Statistical methods in language and linguistic research / G. P. Cantos. – CT: Equinox, 2013. 33. Dabagh R. M. Authorship Attribution and Statistical Text Analysis / R. M. Dabagh // Metodološki zvezki. – 2007. – Vol. 4, № 2. – P. 149–163. 34. Gries Th. S. Quantitative methods in linguistics / Th. S. Gries. – CA: University of California. – 13 p. 35. Mendenhall T. C. The characteristic curves of composition / T. C. Mendenhall // Science. – 1887. – Vol. 9. – P. 237–249.