Ігор КУЛЬЧИЦЬКИЙ

Окремі аспекти квантитативних досліджень української мови

Ігор КУЛЬЧИЦЬКИЙ *– кандидат технічних наук, доцент катедри прикладної лінгвістики Національного університету «Львівська політехніка». Коло зацікавлень: корпусна лінгвістика, комп'ютерна лексикографія, тексти українською мовою, що видані в Західній Україні до 1939–44 років. Електронна адреса: bis.kim@gmail.com*

Статтю присвячено окремим аспектам статистичних досліджень. Коротко описано необхідність використання таких методів у лінгвістиці та охарактеризовано основні напрями їхнього застосування, до яких належать: різноманітні обчислення кількости тих чи тих лінгвальних явищ; побудова на основі обчислених кількісних даних стохастичних моделей мовних явищ; перевірка гіпотез про різні лінгвальні явища статистичними методами; впровадження одержаних результатів у різні галузі, що пов'язані з використанням та вивченням природних мов. Зазначено, що кожну лінгвальну одиницю досліджують як компоненту мовної системи. Зокрема, це її будова та належність до граматичного класу, лексична та синтаксична сполучуваності, місце одиниці в мовній системі і її підсистемах та накладені на неї мовною системою обмеження (наприклад, відсутність деяких словозмінних форм іменників, відсутність ступенів порівняння прикметників і прислівників, неповна дієслівна парадигма і под)., семантичні особливості тощо. Підкреслено, що проведення нових досліджень над наявними та новими збірками текстів дає змогу як підтвердити раніше виявлені, так і визначити нові статистичні параметри та закономірності будови текстів, структурних особливостей різних мов. Дослідження такого типу є основою статистичної типології текстів та мов. Виокремлено такий напрям статистичних досліджень, як атрибуція текстів (встановлення авторства текстів). Подано результати окремих статистичних досліджень творів Василя Стефаника, Михайла Яцкова, Ліни Костенко та Петра Кармансього, які здійснено на кафедрі прикладної лінгвістики Національного університету «Львів-

ська політехніка». Для вказаних творів обчислено частотність символів, їхню милозвучність, визначено можливий мінімальний обсяг вибірки щодо обсягу всіх текстів автора. Стверджено, що таку вибірку необхідно формувати випадковим вибором фраґментів тексту з усіх творів автора. Перевірено й одержано неґативний результат стосовно ґіпотези про те, що частотність символів у творах автора може бути індикатором авторства. Для творів В. Стефаника та М. Яцкова статистично опрацьовано кількість абзаців у творах, довжину абзаців у реченнях, довжину речень у словоформах та довжину словоформ. Зроблено висновки.

Ключові слова: лінгвостатистика, українська мова, В. Стефаник, П. Карманський, частотність, квантитативні дослідження.

Свого часу лорд Кельвін писав, що ми лише тоді щось знаємо про те, що говоримо, коли можемо це виміряти й подати в числах, – інакше наше знання недостатнє й незадовільне (цитуємо за працею Лотфі Заде[1]). На підтвердження цієї тези сучасна лінгвістика, яка багата своїми дослідницькими методами та способами аналізу тексту, тяжіє до застосування квантитативних методів у вивченні лінгвальних явищ. Редактори колективної монографії[2] вказують на такі аспекти лінгвостатистичних досліджень:

– квантитизація за допомогою операціоналізації та вимірювання мовних сутностей та якостей для опису їх кількісними характеристиками;

– кількісний аналіз та опис лінгвістичних і текстових об'єктів;

– числова класифікація лінгвістичних та текстових об'єктів для подальших досліджень або з практичних міркувань;

– розроблення та застосування статистичних процедур для порівняння мовних і текстових об'єктів;

– моделювання лінгвістичних структур, функцій та процесів за допомогою квантитативних моделей та математичних методів;

– побудова теорії на основі пошуку універсальних законів мови й тексту;

– пояснення мовних явищ (властивостей, структур, процесів) засобами теорії;

– введення лінгвістики в загальну систему наук, тобто встановлення відповідностей вивчення міждисциплінарних відносин у формі узагальнення, аналогії або специфікації;

– розроблення ориґінальної лінгвістичної методології щодо конкретних характеристик предмета лінгвістичного дослідження;

– практичне застосування в різних галузях, зокрема в контексті навчання та викладання, психології / психолінгвістики / психіятрії, стилістики /

[1] Заде Лотфи, *Понятие лингвистической переменной и его применение к принятию приближенных решений* (Москва : Мир, 1976), 9.

[2] Reinhard Köhler, Gabriel Altmann, Rajmund Piotrowski, eds. *Quantitative Linguistik : ein internationals Handbuch = Quantitative linquistics : an international handbook* (Berlin, New York: de Gruyter, 2005)

криміналістики, комп'ютерної лінгвістики та мовних технологій, документознавства, контент-аналізу, мовного планування, дослідження масової комунікації тощо.

Ці аспекти умовно можна згрупувати за такими трьома напрямами:

– різноманітні обчислення кількости тих чи тих лінгвальних явищ;

– побудова на основі підрахованих кількісних даних стохастичних моделей мовних явищ;

– перевірка гіпотез про різні лінгвальні явища статистичними методами;

– впровадження одержаних результатів у різні галузі, що пов'язані з використанням та вивченням природних мов[3].

Застосування математичних методів, зокрема й статистичних, у мовознавстві започаткували у середині XIX – на початку XX століть праці російського математика українського походження Віктора Буняковського, німецьких науковців Ернста Ферстемана (Ernst Försteman), Фрідріха Кедінґа (Friedrich Kaeding), російського математика Андрія Маркова, у подальшому їх було розвинуто у працях Ґабріеля Альтмана (Gabriel Altmann), Райнгарда Кьолера (Reinhard Köhler) (Німеччина); Петера Ґжибека (Peter Grzybek) (Австрія); Гейзи Віммера (Geiza Wimmer) (Словаччина); Адама Павловскі (Adam Pawłowski), Ядвіґи Самбор (Jadwiga Sambor) (Польща); Юхана Тулдави (Естонія); Раймунда Піотровского, Анатолія Шайкевич (Росія) та інших[4].

1994 року засновано Міжнародну асоціяцію квантитативної лінгвістики (IQLA), основна мета якої сприяти застосуванню математичних та статистичних методів у лінгвістичному моделюванні, аналізі текстів та суміжних галузях[5]. Очолюють її такі науковці[6]:

Арджуна Туцці (Arjuna Tuzzi), університет Падуї, Італія – президент. Наукові інтереси: статистичний аналіз текстових даних; інструменти соціяльних опитувань (формулювання питань, дизайн опитування, помилки, не пов'язані із вибіркою); процеси та методи оцінювання; аналіз виборчих даних; аналіз політико-інституціонального спілкування.

Джордж Мікрос (George Mikros), Каподістрійський університет, Афіни, Греція – віце-президент. Наукові інтереси: комп'ютерна лінгвістика та стилістика, статистичний аналіз лінгвістичних даних, соціолінгвістичні варіяції, інтелектуальний аналіз тексту, атрибуція та верифікація авторства, інтелектуальний аналіз даних, машинне навчання, фонетика.

Герман Мойсль (Hermann Moisl), Університет Ньюкасл, Великобританія – секретар. Наукові інтереси: загальна комп'ютерна та кількісна лінгвістика;

---

[3] І. М. Кульчицький, «Дослідження довжини речення та слова у творах Романа Іваничука», *Вісник Національного університету «Львівська політехніка»: Інформаційні системи та мережі*, 872 (2017): 139–149.

[4] Там само, 143; В. А. Звегинцев, *Очерки по общему языкознанию* (Москва: МГУ, 1962), 113-151.

[5] The International Quantitative Linguistics Association, http://www.iqla.org/

[6] Ibid, http://www.iqla.org/iqla_contact.html

створення мовного корпусу та кластерний аналіз одержаних із корпусу даних; середньовічні англійська та ірландська мови; динамічні лінгвістичні моделі.

Еммеріх Келіх (Emmerich Kelih), Віденський університет, Австрія – скарбник. Наукові інтереси: фонологія слов'янських мов (особливо складів); загальна графологія; мовна економія, частота, морфосинтаксичні стратегії кодування; типологія стандартизованих слов'янських мов; корпусна лінгвістика (особливо паралельні тексти); кількісний аналіз тексту та мови; мінімальні словники.

Радек Чех (Radek Čech), університет Острави, Чеська Республіка – член ради. Наукові інтереси: кількісний аналіз тексту та кількісний синтаксис (валентність, складні синтаксичні мережі).

Рамон Феррер-і-Канчо (Ramon Ferrer-i-Cancho), політехнічний університет Каталонії, Іспанія – член ради. Наукові інтереси: теорія частотності (слів), теорія порядку (слів), пізнавальні процеси у тварин і спілкування між ними, складні мережі, еволюційна біологія, геноми та теорія інформації.

Свен Науман (Sven Naumann), університет Тріра, Німеччина – член ради. Наукові інтереси: машинне навчання; розширене машинне навчання; оброблення природних мов; автоматичне одержання даних (інформації); корпусна лінгвістика; класифікація текстів; маркування позначками; синтаксичний аналіз (розбір).

Реля Вуланович (Relja Vulanovic), Кентський державний університет, США – член ради. Наукові інтереси: числові методи, сингулярні збурення, математика та квантитативна лінгвістика.

Шейла Емблтон (Sheila Embleton), Йоркський університет, Торонто, Канада – представниця Північної та Південної Америки. Наукові інтереси: історична лінгвістика, соціолінгвістика, діялектологія, математичні та статистичні методи в лінгвістиці, ономастика, семіотика Пірса, ґендерні дослідження, діялектометрія з конкретним застосуванням до британських, фінських та румунських діялектів.

Гаруко Санада (Haruko Sanada), університет Рішшьо́, Токіо, Японія – представник Азії. Наукові інтереси: лінгвістика, японське мовознавство.

Почесним президентом асоціації є Ґабріель Альтман (Gabriel Altmann), Німеччина. Наукові інтереси: розроблення гіпотез про лінгвістичні закони, які виходять із теоретичних припущень, сформульованих математично, а потім емпірично досліджених. Низка мовних законів набули своєї нинішньої форми завдяки Ґ. Альтману, серед них – закон про диверсифікацію, який називають законом Менцерата-Альтмана, та закон Піотровського.

Офіційним періодичним виданням асоціації є журнал «Квантитативна лінгвістика»[7], який заснований 1994 й виходить 4 рази на рік. Це своєрідний міжнародний форум для публікування праць із застосування математики і статистики в лінгвістичних дослідженнях. Зокрема, його тематикою є:

[7] The International Quantitative Linguistics Association, http://www.iqla.org/iqla_journal.html

– моделювання всіх аспектів природної мови в межах теоретичної та історичної лінгвістики, а також соціо-, психо- та нейролінгвістики, діялектології;

– практичне застосування математичних і статистичних методів в опрацюванні природної мови, корпусній лінгвістиці, машинному перекладанні та вивченні мови тощо;

– методологічні проблеми лінгвістичного вимірювання, побудови моделі, вибірки та теорії тесту;

– питання будови філософії мови та мовознавства в межах філософії науки.

Усі статті в цьому журналі проходять жорсткий експертний огляд: первинний огляд редактора та рецензування двома анонімними рецензентами.

Засновник і до сьогодні редактор журналу – Райнгард Кьолер (Reinhard Köhler), університет Тріра, Німеччина. Наукові інтереси: кількісні та системні теорії мовознавства; лінгвістична синергетика.

За час існування асоціації під її егідою проведено 10 наукових конференцій. В останній (2016, Трір, Німеччина[8]) взяли участь 77 науковців, які репрезентували 32 університети з 16 країн світу (Австрія – 3, Великобританія – 2, Греція – 1, Іспанія – 1, Італія – 2, Канада – 1, Китай – 5, Німеччина – 2, Польща – 1, Росія – 3, Словаччина – 1, США – 2, Чехія – 3, Швейцарія – 1, Японія – 4), дві академічні установи (Польща, Росія) та одну аналітично-дослідницьку організацію (США).

Статистичні дослідження української мови розпочато у другій половині минулого століття, коли в Інституті мовознавства ім. О. О. Потебні АН УРСР було створено групу структурно-математичної лінгвістики[9]. На сьогоднішній день статистичні дослідження виконують в Українському мовно-інформаційному фонді, Інституті лінгвістики Київського національного університету ім. Тараса Шевченка, Львівському національному університеті ім. Івана Франка, Національному університеті «Львівська політехніка» та ін.[10]

Отже, за останні три десятиліття квантитативна лінгвістика зазнала бурхливого розвитку як у теорії, так і на практиці. Завдяки їй у лінгвістику було впроваджено кількісні методи й моделі, які використовують у природничих та суспільних науках, що сприяє розвитку нових теоретичних поглядів та розв'язанню практичних проблем у її різних галузях.

За Валентиною Перебийніс, кількісні методи уточнюють результати досліджень не тільки в мовленні, але й у мові, уможливлюють науково обґрунтовані, часом непередбачувані висновки[11]. Разом із забезпеченням ві-

---

[8] http://www.iqla.org/qualico2016abstracts.pdf

[9] Соломія Бук, «Статистичні характеристики лексики основних функціональних стилів української мови: спроба порівняння», *Лексикографічний бюлетень*, 13 (2006): 166–170.

[10] В. А. Широков, *Інформаційна теорія лексикографічних систем* (Київ: Довіра, 1998); Соломія Бук, «Сучасні методи дослідження мови письменника у слов'янознавстві», *Проблеми слов'янознавства,* 61(2012): 86–95; Соломія Бук, «Лінгвостатистичний опис «Не спитавши броду» Івана Франка», *Вісник Львівського університету,* 55 (2011): 230–242; Mova.info «Лінгвістичний портал», *http://www.mova.info/*

[11] В. І. Перебийніс, «Що дає статистика лінгвістам?», *Вісник Київського лінгвістичного університету*, VI, 2 (2003): 27.

рогідности результатів статистичні методи уможливлюють розкриття таких властивостей мовних одиниць та будови тексту, які без них неможливо було би виявити. Ефективність квантитативних методів зумовлено низкою причин[12]:

– одержані під час дослідження точні кількісні дані завжди можна перевірити, на відміну від якісних, на кшталт «незначно», «часто» тощо;

– завжди можна визначити – випадково чи істотно коливаються значення показників, які зіставляють;

– можна визначити необхідну та достатню для вірогідних висновків кількість дослідницького матеріялу;

– необхідна формалізація під час визначення одиниць дослідження мінімізує чинник суб'єктивности та забезпечує вірогідні результати високого ступеня точности;

– такі методи дають змогу дійти правильних висновків.

Кожну лінгвальну одиницю досліджують як компоненту мовної системи. Зокрема, це її будова та належність до граматичного класу, лексична й синтаксична сполучуваності, місце одиниці в мовній системі і її підсистемах та накладені на неї мовною системою обмеження (наприклад, відсутність деяких словозмінних форм іменників, відсутність ступенів порівняння прикметників і прислівників, неповна дієслівна парадигма і под)., семантичні особливості тощо[13].

Статистично досліджуючи будову лінгвістичних одиниць, до уваги беруть їхню довжину, морфемну чи словотвірну структуру, кількість складів, морфем або слів, які її утворюють.

У граматичних дослідженнях вивчають розподіл лінгвістичних одиниць за частинами мови, словозмінними класами, типами словозміни тощо. Аналізуючи синтаксичну й лексичну сполучуваність, розглядають моделі синтаксичних конструкцій, колокації тощо.

У семантичних дослідженнях зосереджуються на кількості значень мовних одиниць, лексичних групах, до яких ці одиниці належать, характер їхніх лексичних та граматичних значень.

Оскільки кожна із цих ознак лінгвістичної одиниці може впливати на її частоту, тобто частота відображає ту чи ту її властивість (певну сукупність властивостей), виявити такі взаємозв'язки можна лише за допомогою статистичних обстежень.

Особливості вживання конкретної одиниці в певному тексті визначають її функційні властивості: частота, позиція, сполучуваність, яка залежить від характеру тексту, від функційного чи авторського стилю і змінюється від тексту до тексту.

---

[12] Там само, 29.

[13] Там само, 28.

Щоб виявити відмінність між частотами мовної одиниці в різних стилях та жанрах, встановлюють і закономірності функціювання одиниці в цих стилях, і особливості останніх. Частоти одиниць, істотно різні в різних стилях, називаються статистичними параметрами стилів, на яких базує свої висновки нова галузь мовознавства – стилеметрія[14]. Ці параметри властиві всім рівням мови, оскільки тексти можуть істотно відрізнятися за частотами фонем, морфем, слів і лексичних груп, синтаксичних конструкцій.

Проведення нових досліджень наявних та нових збірок текстів дає змогу як підтвердити раніше виявлені, так і визначити нові статистичні параметри та закономірності будови текстів, структурних особливостей різних мов. Дослідження такого типу є основою статистичної типології текстів та мов.

У прикладному мовознавстві атрибуцію тексту розуміють як вивчення текстового матеріялу, щоб встановити авторство або одержати певну інформацію про особистість автора чи умови, за яких створено текст. Для цього необхідно розв'язати чи ідентифікаційну, чи діягностичну задачу[15].

Розв'язуючи ідентифікаційні задачі, припускають, що автор тексту відомий дослідникові. За допомогою задач цього типу:

– підтверджують або відкидають авторство конкретної особи;
– перевіряють, чи та сама особа написала весь текст;
– перевіряють факт, що справжній автор – той, хто написав текст.

За допомогою діягностичних задач визначають такі характеристики автора, як місце народження та постійного проживання, рідна мова, рівень освіти, знання іноземних мов, а також підтверджують чи спростовують факт свідомого спотворення мови та ін. У такому разі припускають, що автор тексту не відомий, тому порівняти текст, що досліджують, з авторським неможливо.

Методи атрибуції застосовують на пунктуаційному, ортографічному, синтаксичному, лексико-фразеологічному та стилістичному рівнях.

На пунктуаційному рівні з'ясовують особливості вживання пунктуаційних знаків, специфічні стосовно останніх помилки автора і т. ін.

На ортографічному рівні виявляють специфічні помилки у правописі.

На синтаксичному рівні приділяють увагу типовим синтаксичним структурам, перевазі розповідних, питальних чи окличних речень, уживанню активного чи пасивного стану, порядку слів у реченні, специфічним синтаксичним помилкам тощо.

На лексико-фразеологічному рівні визначають кількісний та якісний словниковий запас автора, зокрема характерні особливості вживання фразеологізмів, схильність до використання рідковживаних чи іншомовних слів, діялектизмів, архаїзмів, неологізмів, професійних термінів, арґотизмів, навички вживання прислів'їв, приказок, афоризмів і т. ін.

---

[14] Валентина Перебийніс, ред., *Статистичні параметри стилів* (Київ: Наукова думка, 1967).

[15] Т. В. Батура, «Формальные методы определения авторства текстов», *Вестник НГУ*, 10, 4: 81–94.

На стилістичному рівні визначають жанр тексту, його загальну будову, зокрема для літературних творів – сюжет, типові образотворчі прийоми (метафора, іронія, алегорія, гіпербола, порівняння), стилістичні фігури (антитеза, риторичне питання тощо), інші характерні мовні прийоми.

До авторського стилю зазвичай зараховують синтаксичний, лексико-фразеологічний та стилістичний рівні. Їхній аналіз є доволі складний і водночас збуджує найбільший інтерес.

Методи аналізу стилю ділять на дві групи: експертні й формальні. В експертних методах текст опрацьовують професійні лінгвісти – експерти. Формальні методи ділять на методи, що базовані на машинному навчанні (баєсівський класифікатор, нейронні мережі, дерева рішень, метод опорних векторів, генетичні алгоритми, метод k-найближчих сусідів) та методи, що базовані на статистичних дослідженнях (одновимірні: критерій Стьюдента, двосторонній критерій Фішера, хі-квадрат Пірсона; багатовимірні: метод головних компонент, критерій Колмогорова–Смірнова, ланцюги Маркова, хі-квадрат Пірсона для розподілів, статистичний кластерний аналіз тощо). Базу формальних методів зазвичай становлять порівняння обчислених характеристик текстів. Як формальну модель текст чи тексти автора зображують вектором параметрів, кожний із яких об'єктивно розкриває ту чи ту характеристику тексту[16].

Для порівняння двох текстів беруть інтегральну характеристику, яку обчислюють тим чи тим способом (наприклад, ентропію)[17]. У найпростіших ситуаціях сукупність параметрів розглядають як звичайний вектор у n-мірному декартовому просторі, а за інтеґральну характеристику беруть звичайну декартову відстань між кінцями відповідних їм векторів. Здебільшого за параметри, що характеризують текст, беруть його ті чи ті статистичні характеристики: частоту вживання окремих частин мови, знаків пунктуації, конкретних слів, фразеологізмів, архаїзмів, рідковживаних та іншомовних слів, кількість і довжину речень (виміряну у словах, складах, символах), обсяги словника, кількість повнозначних і службових слів, середню довжину речення, відношення кількости дієслів до загальної кількости слововживань у тексті тощо.

Вибираючи параметри, треба враховувати, що не всі вони придатні для атрибуції текстів через один із двох таких недоліків:

– Відсутність рівноваги. Розкид значень параметра в текстах одного автора такий великий, що діяпазони його значень і різних авторів перекриваються. Очевидно, що такий параметр не дасть змоги розрізнити авторство, а його використання у складі групи створюватиме додатковий побічний шум.

– Відсутність розрізнювальної здатности. Значення параметра визначають властивостями мови, якою написані тексти, а не індивідуальними

[16] Там само, 83.

[17] Там само, 84.

особливостями мови автора. Такий параметр набуває близьких значень для будь-яких текстів будь-якого автора.

З огляду на це параметри перед використанням належить досліджувати на стійкість та розрізнювальну здатність. При цьому бажано використовувати максимально можливу кількість авторів та виходити із припущення, що формальний параметр має відповідати таким умовам[18]:

Масовість. Параметр повинен збігатися з тими характеристиками тексту, які автор лише незначною мірою контролює на свідомому рівні. Це необхідно для того, щоби відкинути можливість свідомого спотворення чи зміни автором типового для нього стилю або стилізації під іншого автора.

Стійкість. Параметр передбачає збереження постійного значення для одного автора. Очевидно, що внаслідок випадкових чинників обов'язково буде певне відхилення значень від середнього значення, однак воно має бути незначним.

Розрізнювальна здатність. В ідеалі параметр має набувати істотно різних значень (незначні коливання можливі для одного автора) для різних авторів. Варто зазначити, що вибрати параметри, які ґарантовано розрізняють двох авторів, дуже складно. Хай би які не були параметри, завжди є ймовірність того, що за цими параметрами два або більше авторів виявляться близькими через випадковий збіг. Тому, як свідчить практика, достатньо, щоб параметр давав змогу переконливо розрізняти між собою різні групи авторів, тобто щоби була досить велика кількість груп авторів, для яких середнє значення параметра значно відрізняється. У такому разі параметр, мабуть, не дасть змоги розрізнити текстів авторів однієї групи, однак уможливить розрізнення текстів авторів, що перебувають у різних групах.

Застосування статистичних методів у лінгвістичних дослідженнях покажемо на прикладах визначення складників статистичного профілю творів окремих авторів. З усього різноманіття таких характеристик зупинимося на частотності графем, довжині абзаців, речень та словоформ. Такі дослідження проводять на кафедрі прикладної лінгвістики Національного університету «Львівська політехніка».

Дослідницький матеріял сформовано так. Для дослідження вибрано твори Василя Стефаника[19], Михайла Яцкова[20], Ліни Костенко[21] та Петра Карманського[22]. Наявність у матеріялі творчого доробку двох новелістів, які творили в різний час, та двох поетів, один із яких репрезентує Наддні-

---

[18] Ігор Кульчицький, «Ентропія одно- та двограм символіввукраїномовних текстах», *Науковий вісник Волинського національного університету імені Л. Українки*, 6, 2016: 183–189.

[19] Василь Стефаник, «Межа», *Літературно-науковий вісник*, 92, 2 (1927): 97–98; Василь Стефаник, «Портрет», *Твори* (Харків: ДВУ, 1929, 94–95); Василь Стефаник, *Твори* (Львів: Видавнича Спілка «Діло», 1933); Василь Стефаник, «Шкільник», *Рідна школа*, 1932: 2–4.

[20] Михайло Яцків, *Вибрані твори* (Київ: Дніпро, 1973).

[21] Ліна Костенко, *Вибране* (Київ: Дніпро, 1989).

[22] П. С. Карманський, *Поезії* (Київ: Укр. письменник, 1992).

прянську, а другий – Наддністрянську Україну, дала змогу порівняти твори за визначеними показниками як у стильовому, так і в часовому та реґіональному розрізах. Для проведення досліджень твори названих авторів перетворено на електронну форму та звірено з паперовими ориґіналами. Отже, дослідницька збірка містить: усі (55) відомі на сьогодні новели В. Стефаника, вибрані (42) новели М. Яцкова, 328 поезій Л. Костенко та 320 поезій П. Кармансько́го.

Для проведення статистичних досліджень на рівні символів (графем) із творів кожного автора утворено масив символьних рядків за такими правилами:

– у текстах залишено тільки символи розширеної української абетки, до якої залучено символи традиційної абетки, пропуск, дефіс та апостроф;

– усі букви перетворено на великі;

– між словами залишено лише один пропуск;

– утворено один суцільний текст послідовним приєднанням через пропуск текстів: для новел в алфавітному порядку за назвами, для поезії – за порядком розміщення у збірці;

– утворений текст поділено на масив символьних рядків однакової довжини, яка становить 108 символів (трикратний розмір розширеного алфавіту – вибрано автором інтроспективно); якщо останній рядок мав довжину меншу ніж 108 символів, то його відкинуто.

Для досліджень довжин абзаців, речень та словоформ у новелах В. Стефаника та М. Яцкова позначено початок і кінець речень та абзаців.

Методика обчислень була такою.

Усі статистичні показники обчислювали за стандартними для математичної статистики формулами[23]. Зокрема, середнє обчислювали за формулою:

$$\bar{X} = \frac{\sum x_i n_i}{\sum n_i},$$

де x_i – варіанта, n_i – кількість наявностей варіанти в дослідах, i – номер варіанти, а середнє квадратичне відхилення – за формулою:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{\sum n_i}},$$

де x_i – варіанта, n_i – кількість навяностей варіанти в дослідах, i – номер варіанти, $\bar{x}$ – середнє значення.

Міру коливання середньої частоти обсилювали за формулою:

[23] Валентина Перебийніс, ред., *Статистичні параметри стилів,* 23–43.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{\sum n_i}},$$

а стандартну похибку – за формулою:

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{\sum n_i - 1}},$$

де σ – середнє квадратичне відхилення, n_i – кількість наявностей варіанти в дослідах, і – номер варіанти, $\bar{x}$ – середнє значення.

Відносну похибку обчислювали за формулою:

$$\varepsilon = \frac{1{,}96\sigma_{\bar{x}}}{\bar{x}},$$

де $\bar{x}$ – середнє, $\sigma_{\bar{x}}$ – міра коливання середньої частоти.

Для кожного масиву виконано такі дії:

Крок 1. Обчислено абсолютну (кількість) та відносну частоту кожного символу розширеної української абетки в цілому символьному масиві. За одержаними частотами для кожного символу визначено його ранґ та обчислено ентропію за формулою[24]:

$$H = -\sum_{i=1}^{36} p_i \log_2 p_i,$$

де p_i – відносна частота і-го символу.

Крок 2. Визначено розмір сеґмента тексту в рядках. Розмір вибірки встановлено діленням націло кількости рядків у масиві на 100. Залишок використано для організації послідовного вибору сеґментів.

Крок 3. Послідовно, починаючи з першого рядка, з символьного масиву вибирали сеґмент визначеної в попередньому кроці довжини. Для кожного символу розширеної української абетки у вибраному сеґменті визначали абсолютну й відносну частоти, ранґ та ентропію. Після досягнення кінця масиву вибір із подальшими обчисленнями послідовних сеґментів починали з другого рядка, потім – з третього і т. д. Кількість зміщень від початку масиву була регламентована залишком від ділення націло кількости рядків у масиві на 100. Після завершення вибору послідовних сеґментів визначено середню частоту та ентропію кожного символу.

Крок 4. Аналогічні обчислення проведено для сеґментів, які формували з рядків масиву за допомогою генератора випадкових чисел. Розмір сеґмента та кількість вибірок відповідала значенням попереднього кроку.

[24] А. М. Яглом, И. М. Яглом, *Вероятность и информация* (Москва: КомКнига, 2007).

Крок 5. За результатами проведених обчислень проаналізовано відповідність частот символів у цілому масиві та вибраних сеґментах. За теоретичне підґрунтя взято критерій узгодженості К. Пірсона (χ^2)[25]. За гіпотетичну теоретичну функцію розподілу прийнято частотний розподіл символів у дослідницькому масиві тексту. Для кожного вибраного сеґмента обчислювали статистику критерію χ^2_{exp}. За нульову гіпотезу H_0 прийняли твердження: «у тексті-вибірці розподіл частот символів розширеної української абетки не відрізняється від відповідного розподілу в текстовому масиві». $t_{cr} = \chi^2_{1-\alpha,k-1}$ визначали за стандартною таблицею для рівня значущости α=0,05 та відповідному ступені свободи k-1. Число k (максимально – кількість символів розширеної української абетки, дорівнює 36) залежало від приведення одержаних послідовностей до вимоги рівності мінімального значення в послідовності не менше ніж 5. Якщо отримували, що $\chi^2_{\text{exp}} \geq t_{cr}$, то гіпотезу відхиляли, інакше її приймали.

Крок 6. За одержаними результатами визначено усереднений ранґ частоти кожного символа. Таким ставало те значення ранґу, яке символ займав найбільшу кількість разів як у всьому тексті, так і в кожному сеґменті. Якщо ранґи двох символів збігалися, то враховувалося абсолютне значення кількости.

Крок 7. Для всього символьного масиву та окремо для послідовно й випадково вибраних його сеґментів обчислено розподіл символів за типами та милозвучність тексту. Милозвучність визначали як відсоток сукупности голосних, сонорних та дзвінких букв.

Крок 8. Для кожного масиву проведено аналіз розміру сеґмента, починаючи з якого його частота символів відповідала частоті символів всього масиву. З цією метою початковий розмір сеґмента встановили в одну соту розміру масиву. Для всіх можливих сеґментів такої довжини обчислювали за критерієм узгодженості Пірсона (див. попередній крок) кількість збігів частот символів у сеґменті й масиві. Після того розмір масиву збільшували на одну соту й операцію повторювали. Дослід закінчували тоді, коли розмір сеґмента дорівнював приблизно 99 % розміру всього масиву.

На завершальному етапі пораховано довжини абзаців, речень та словоформ у новелах В. Стефаника та М. Яцкова.

Усі обчислення виконано за допомогою власних програм, написаних мовою Python.

Внаслідок обчислень одержано такі показники.

Кількість символів у масивах рядків, утворених із текстів кожного автора, подано в таблиці 1.

[25] Ружевич Н. А. *Математична статистика* (Львів: Видавництво Національного ун-ту Львівська політехніка, 2001).

Таблиця 1. *Кількість символів у масивах рядків, що утворені з текстів автора*

Автор	**Кількість символів**
Василь Стефаник	286 092
Михайло Яцків	246 888
Ліна Костенко	237 384
Петро Карманський	218 700

Як бачимо, кількість символів у прозових творах та в поезії зіставна.

У таблиці 2 подано частотність символів як у цілих масивах (колонка **Т**), так і в сеґментах (колонка **С**, значення усереднено) творів автора. У списку символів «<SP>» означає «пропуск».

Таблиця 2. *Частотність символів у творах авторів*

Символ	**Частота**							
	В. Стефаник		**М. Яцків**		**Л. Костенко**		**П. Карманський**	
	Т	**С**	**Т**	**С**	**Т**	**С**	**Т**	**С**
А	0,0811	0,0811	0,0745	0,0746	0,0662	0,0662	0,0616	0,0615
Б	0,0205	0,0204	0,0164	0,0163	0,0164	0,0165	0,0162	0,0163
В	0,0429	0,0431	0,0474	0,0476	0,0423	0,0423	0,0426	0,0426
Г	0,0139	0,0139	0,0147	0,0147	0,0133	0,0132	0,0140	0,0140
Ґ	0,0007	0,0007	0,0002	0,0002	0,0001	0,0001	0	0
Д	0,0293	0,0293	0,0286	0,0286	0,0256	0,0255	0,0261	0,0262
Е	0,0393	0,0393	0,0389	0,0389	0,0446	0,0447	0,0396	0,0397
Є	0,0052	0,0052	0,0041	0,0041	0,0035	0,0035	0,0042	0,0042
Ж	0,0082	0,0082	0,0072	0,0072	0,0101	0,0101	0,0088	0,0088
З	0,0177	0,0178	0,0183	0,0183	0,0166	0,0166	0,0193	0,0193
И	0,0607	0,0607	0,0540	0,0540	0,0540	0,0541	0,0541	0,0542
І	0,0436	0,0437	0,0475	0,0475	0,0505	0,0507	0,0521	0,0521
Ї	0,0043	0,0044	0,0049	0,0049	0,0040	0,0040	0,0034	0,0034
Й	0,0129	0,0129	0,0108	0,0108	0,0114	0,0114	0,0137	0,0137
К	0,0318	0,0319	0,0309	0,0309	0,0297	0,0297	0,0256	0,0256
Л	0,0328	0,0328	0,0361	0,0361	0,0346	0,0345	0,0332	0,0331
М	0,0256	0,0256	0,0242	0,0241	0,0273	0,0274	0,0306	0,0307
Н	0,0424	0,0424	0,0474	0,0472	0,0490	0,0490	0,0502	0,0503

Символ	Частота							
	В. Стефаник		М. Яцків		Л. Костенко		П. Карманський	
	Т	С	Т	С	Т	С	Т	С
О	0,0744	0,0744	0,0752	0,0753	0,0694	0,0692	0,0639	0,0640
П	0,0213	0,0214	0,0243	0,0243	0,0217	0,0217	0,0211	0,0211
Р	0,0299	0,0299	0,0345	0,0345	0,0367	0,0367	0,0382	0,0383
С	0,0320	0,0320	0,0343	0,0343	0,0354	0,0355	0,0388	0,0388
Т	0,0454	0,0453	0,0412	0,0412	0,0443	0,0444	0,0435	0,0434
У	0,0273	0,0273	0,0289	0,0289	0,0292	0,0292	0,0286	0,0287
Ф	0,0006	0,0006	0,0009	0,0009	0,0012	0,0013	0,0006	0,0006
Х	0,0096	0,0096	0,0107	0,0107	0,0109	0,0109	0,0123	0,0123
Ц	0,0053	0,0053	0,0053	0,0054	0,0064	0,0064	0,0046	0,0046
Ч	0,0086	0,0086	0,0109	0,0109	0,0115	0,0115	0,0103	0,0103
Ш	0,0086	0,0086	0,0076	0,0077	0,0073	0,0074	0,0072	0,0072
Щ	0,0035	0,0035	0,0042	0,0042	0,0049	0,0050	0,0047	0,0047
Ь	0,0084	0,0085	0,0125	0,0126	0,0165	0,0165	0,0171	0,0171
Ю	0,0063	0,0063	0,0071	0,0071	0,0067	0,0067	0,0092	0,0092
Я	0,0162	0,0162	0,0203	0,0204	0,0190	0,0189	0,0198	0,0198
’	0,0001	0,0001	0,0010	0,0009	0,0014	0,0014	0,0014	0,0014
-	0,0009	0,0009	0,0009	0,0009	0,0010	0,0010	0,0012	0,0011
<SP>	0,1887	0,1888	0,1742	0,1743	0,1775	0,1775	0,1822	0,1822

Для кожного автора частоти символів у всьому тексті та в його сеґментах порівняли за критерієм узгодженості Пірсона. Порівняння засвідчило, що частоти символів належать до однієї генеральної сукупності. Отже, доходимо висновку, що для визначення частотности символів у творах одного автора не потрібні всі його твори, а достатньо вибірки з текстів його творів. Спосіб одержання та розмір вибірки буде обговорено в подальшому викладі матеріялу.

Зовсім інші, дещо несподівані для автора, результати дало аналогічне порівняння частотности символів творів досліджуваних авторів. Додатково для цього утворено масив рядків символів із творів усіх авторів та використано результати частотности символів, що одержані на масиві художніх творів українською мовою обсягом близько 12 мільйонів символів в Українському мовно-інформаційному фонді НАНУ[26]. Результати обчислення χ^2_{exp} подано в таблиці 6.

[26] В. А. Широков, *Інформаційна теорія лексикографічних систем.*

Таблиця 3. *Значення* χ^2_{exp} *для критерію узгодженості Пірсона в текстових масивах*

	В. Стефаник	**М. Яцків**	**Л. Костенко**	**П. Карманський**	**Усі разом**	**УМІФ**
В. Стефаник		4 915,28	9 732,77	10 619,12	15 999,91	595 609,76
М. Яцків	4 108,43		2 001,57	3 269,28	2 844,06	336 053,40
Л. Костенко	7 560,72	1 951,62		1 415,46	4 100,14	464 173,59
П. Карманський	39 582,89	4 670,69	1 807,84		21 765,88	438 749,87
Усі разом	2 765,80	696,24	976,60	1 775,66		353 203,86
УМІФ	7 376,82	4 282,76	7 158,29	7 808,78	19 287,34	

За максимально допустимого $t_{cr} = 49$ можна стверджувати, що частота символів у творах кожного автора суворо індивідуальна й не збігається ні з частотою символів у об'єднаному масиві текстів, ні з частотою символів, одержаною в УМІФ. Причину такого явища, на мою думку, необхідно шукати в подальших дослідженнях, збільшуючи як кількість творів, так і кількість авторів.

Такі результати породжують припущення, що частотність символів у творах автора може слугувати розрізнювальною ознакою під час встановлення авторства тексту. Для перевірки цього припущення, обчислюючи частоти символів у тексті чи його сеґменті, визначали ентропію символу, вважаючи, що саме вона й буде ознакою-розрізнювачем. Узагальнені результати подано в таблиці 4.

Аналіз значень ентропії символів для текстів творів авторів, залучених до дослідження, та сеґментів текстів доводить, що вона не може бути показником належности твору конкретному автору, оскільки її значення між авторами перекриваються. Підкреслимо, що йдеться тільки про твори 4 авторів. Загальних висновків можна буде дійти лише за значного збільшення кількости авторів та їхніх творів. Окрім того, доцільним, на мою думку, був би пошук інших розрізнювальних індикаторів авторства, базованих не тільки на частотності символів, але й на частотності комбінацій двох, трьох і т. ін. символів. Хоча проведені дослідження з символьними двограмами[27] позитивного результату не дали.

[27] Ігор Кульчицький, «Ентропія одно- та двограм символів в україномовних текстах», *Науковий вісник Волинського національного університету імені Л. Українки*, 6, 2016: 183–189.

Таблиця 4. *Ентропія символів у текстах авторів*

Автор	Увесь текст	Вибірка			
		Тип	Ентропія		
	Ентропія		Мінімальна	Середня	Максимальна
В. Стефаник	4,40874	послідовна	4,3009	4,389	4,4862
		випадкова	4,3236	4,3992	4,4623
М. Яцків	4,46697	послідовна	4,3536	4,4468	4,5439
		випадкова	4,3803	4,456	4,541
Л. Костенко	4,48014	послідовна	4,3448	4,4609	4,5414
		випадкова	4,3868	4,4677	4,5503
П. Кармaнський	4,47987	послідовна	4,3633	4,4576	4,5373
		випадкова	4,3865	4,4663	4,5545

Одержані результати заперечують і припущення про стабільність інваріантности частотности букв у текстах кожної мови, зокрема української. На рис. 1 показано частотні ранґи символів, які вони одержували у проведених обчисленнях частотности як у кожному з масивів, так і у кожному з його сеґментів.

Темно-жовтим кольором позначено нормований ранґ кожного символу, жовтим – ранґи, частота одержання яких близька до нормованого, зеленим – решта ранґів із меншими кількісними значеннями (зокрема з 1-2), які було присвоєно символу. Проведене дослідження дало змогу виокремити групу найчастотніших (таблиця 5), середньочастотних (таблиця 6) та низькочастотних (таблиця 7) символів у текстах творів названих письменників. Подальші дослідження уможливлять уточнення цих показників для української мови.

Таблиця 5. *Найчастотніші символи української мови*

Символ	Кількість разів одержання нормованого ранґу	Ранґ				
		нормований	Розмах		Найбільш можливі	
			мінімум	максимум	мінімум	максимум
<SP>	52 404	1	1	1	1	1
О	26 780	2	2	10	2	3
А	24 559	3	2	9	2	3
И	29 356	4	2	11	4	5
І	13 239	5	2	14	4	7

Символ	Кількість разів одержання нормованого рангу	Ранґ				
		нормований	Розмах		Найбільш можливі	
			мінімум	максимум	мінімум	максимум
Н	13 101	6	2	15	5	7
В	9 812	7	3	18	5	9
Т	10 814	8	3	18	5	9
Е	10 830	9	2	17	7	10
Р	10 342	10	4	19	10	12
С	11 854	11	3	19	10	12
Л	9 478	12	4	13	10	12

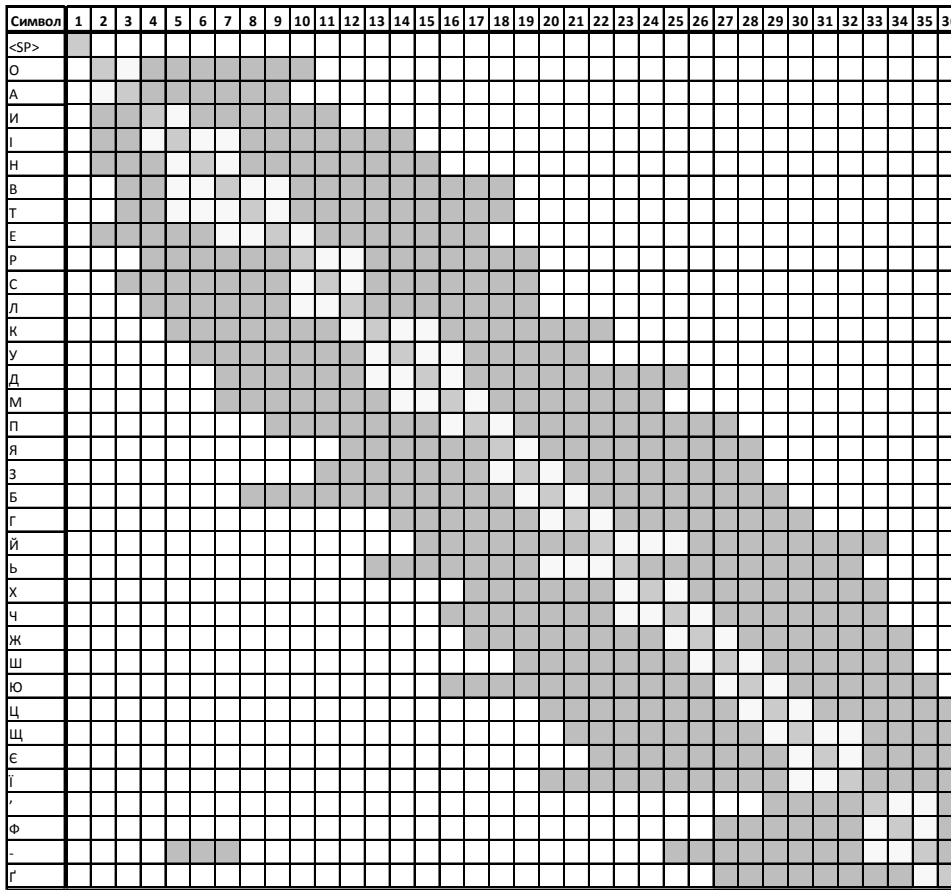

Рис. 1. *Ранґи частотности символів у творах В. Стефаника, М. Яцкова, Л. Костенко та П. Карманського (Кольоровий онлайн).*

Нормований ранґ – це ранґ, який символ отримував найбільшу кількість разів у всіх (54 404) обчисленнях частоти символів. Якщо ранґ збігався для декількох символів, то їх ранжували за спаданням значення кількости. Розмах – це найбільше та найменше значення ранґу, яке отримував символ хоча б один раз. Зрозуміло: що значення ранґу менше, то сам ранґ вищий. Найбільш можливі – це найближчі до нормованого ранґи, які одержував символ.

Таблиця 6. *Середньочастотні символи української мови*

Символ	Кількість разів отримання нормованого ранґу	Ранґ				
		Нормований	Розмах		Найбільш можливі	
			мінімум	максимум	мінімум	максимум
К	10 032	13	5	22	12	15
У	11 642	14	6	21	13	16
Д	11 572	15	7	25	13	16
М	10 883	16	7	24	14	17
П	15 778	17	9	27	16	18
Я	13 353	18	12	28	18	19
З	13 883	19	11	28	18	20
Б	11 230	20	8	29	19	21
Г	12 107	21	14	30	20	22
Й	8 631	22	15	33	22	25
Ь	7 620	23	13	32	20	23
Х	10 074	24	33	33	23	25
Ч	9 685	25	16	33	23	26
Ж	9 560	26	17	34	25	27
Ш	10 374	27	18	34	26	28
Ю	9 540	28	16	35	27	29

Таблиця 7. *Низькочастотні символи української мови*

Символ	Кількість разів одержання нормованого ранґу	Ранґ				
		Нормований	Розмах		Найбільш можливі	
			мінімум	максимум	мінімум	максимум
Ц	10 735	29	20	36	28	30
Щ	11 724	30	21	36	29	32
Є	13 170	31	22	36	30	32

Символ	Кількість разів одержання нормованого рангу	Ранґ				
		Нормований	Розмах		Найбільш можливі	
			мінімум	максимум	мінімум	максимум
Ї	12 824	32	20	36	30	32
’	15 694	33	29	36	33	35
Ф	14 977	34	27	36	33	35
-	17 839	5	25	36	33	35
Ґ	32 391	36	27	36	35	36

Ще один досліджуваний показник – це милозвучність текстів. Розподіл символів за типами подано в таблиці 8.

Обчислення милозвучности дали такі результати:

– Василь Стефаник – 78,7 %;
– Михайло Яцків – 78,5 %;
– Ліна Костенко – 78,0 %;
– Петро Карманський – 78,3 %.

Таблиця 8. *Розподіл символів за типами*

	В. Стефаник	М. Яцків	Л. Костенко	П. Карманський
Голосні	32,60 %	31,90 %	31,40 %	30,00 %
Приголосні:	44,00 %	45,20 %	45,20 %	45,70 %
сонорні	18,66 %	20,07 %	20,16 %	20,84 %
дзвінкі	9,02 %	8,54 %	8,23 %	8,45 %
глухі	16,32 %	16,63 %	16,86 %	16,41 %
Допоміжні	23,40 %	22,90 %	23,40 %	24,30 %

До допоміжних символів належать ті з 36 символів, які не віднесені ні до голосних, ні до приголосних.

Порівняння одержаних величин із показниками, поданими для українськомовних текстів у монографії «Статистичні параметри стилів», показало, що милозвучність творів Василя Стефаника та Михайла Яцкова найближча до милозвучности творів суспільно-політичного (78,7 %), Ліни Костенко та Петра Карманського – до драматургічного (77,9 %) стилів. На мою думку, видимої причини такого явища ще вказати не можна. Цілком можливо, що проблема полягає в доборі текстової інформації, але поданий у монографії матеріял дослідження був дібраний коректно. На жаль, за неофіційною інформацією, первинні матеріали цього дослідження втрачені, тому відтворити їх неможливо. Відповідь, мабуть, за майбутніми дослідженнями.

Ще одне дослідження проведено для визначення мінімального розміру вибірки тексту, після якого частотність символів збігається з частотністю символів у всіх творах автора. Як було сказано вище, частоту символів порівнювали з частотою у всьому тексті, для всіх можливих сеґментів певного розміру, починаючи від сеґмента розміром приблизно в одну соту розміру тексту з поступовим збільшенням на таку ж величину. Сеґменти утворено як послідовним, так і випадковим добором рядків символів фіксованої довжини. Для порівняння результатів під час послідовного й випадкового вибору рядків символів у масиві текстів кількість сеґментів в обох випадках робили однаковою. Фіксували кількість збігів за критерієм узгодженості Пірсона. Результати для творів Василя Стефаника поданона рис. 2. На горизонтальній осі відкладено розмір сеґмента у відсотках до розміру текстового масиву. На вертикальній осі вказано кількість збігів частотности символів у сеґменті та всьому масиві за критерієм узгодженості Пірсона.

Перший факт, який впадає в око: випадковий вибір рядків тексту для формування вибірки дає збіг у понад 90 % випадках уже за розміру останньої в 1 % від розміру масиву текстів, збіг понад 99 % – за досягнення розміру 22,6 %, а 100 % – за 40,2 %. У творах Михайла Яцкова такі показники збігів дають розміри вибірки в 1 %, 22,1 %, 42,3 % від обсягу текстового масиву, у творах Ліни Костенко – 7,6 %, 26,8 %, 43 %. За послідовного вибору рядків для формування вибірки у творах Василя Стефаника теж спостерігаємо цікаве явище — за розміру вибірки від 1 % до 10,8 % кількість збігів частотности падає від 12,5 % до 0 %, далі за розмірів від 11,8 % до 18,7 % кількість збігів нульова, а між розмірами від 19,6 % до 47,1 кількість збігів має хаотичний характер. Від розміру вибірки у 48,1 % кількість збігів стабільно зростає й за розміру 74,6 % стає стовідсотковою. Пояснення таких фактів потребує окремих досліджень. Аналогічні результати одержані для творів Михайла Яцкова та Ліни Костенко.

Абсолютно іншу картину спостерігаємо для творів П. Карманського. У творах цього автора за умови формування вибірки з рядків, вибраних у випадковий спосіб, за розмірів від 1 % до 73,1 % кількість збігів хоча й більша ніж 58%, але має хаотичний характер, а 100% збігів настає за 74,1%. За формування вибірки послідовним вибором рядків хаотичність кількости збігів ще більша, а стовідсотковий збіг настає за розміру вибірки у 87,9 %. Частково такі результати можна пояснити тим, що дослідницька збірка містить твори різних років, але не у пропорційному співвідношенні. Однак без додаткових досліджень одержати вірогідну відповідь не можна.

З одержаних результатів випливає, що для одержання вірогідної частоти символів у творах окремого автора вибірку належить формувати випадковим добром частин тексту з різних творів. Мінімальний розмір вибірки має становити хоча би 25 % відсотків від обсягу всіх творів у символах, а 44 % дають практично достовірний результат. Хоча подальші дослідження можуть скоригувати ці цифри.

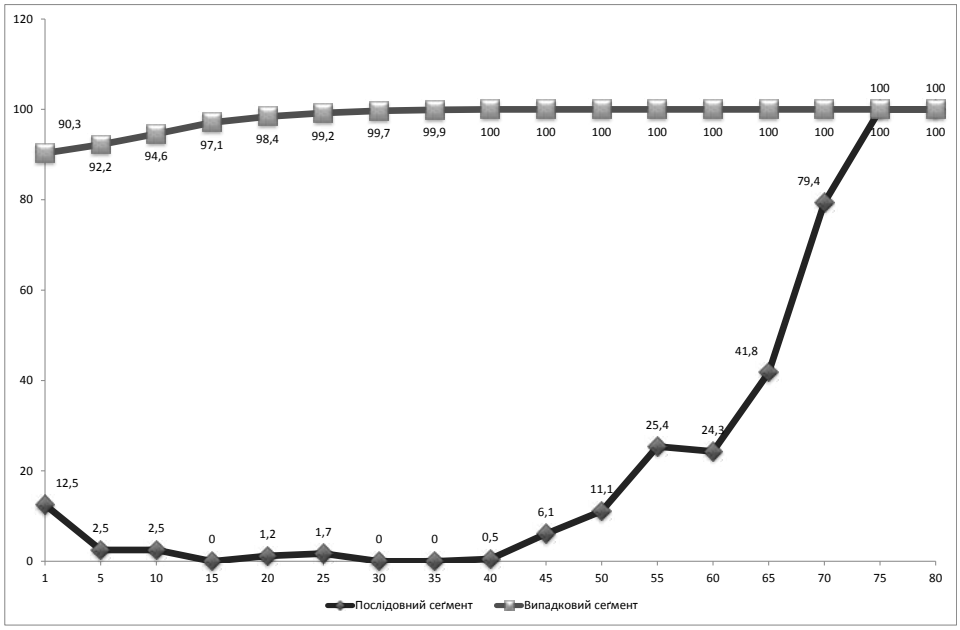

Рис. 2. *Збіг частотності символів у сегментах та повному текстовому масиві творів В. Стефаника*

Низку статистичних показників одержано для абзаців, речень та словоформ у творах Василя Стефаника та Михайла Яцкова. Результати подано в таблицях 9–12.

Таблиця 9. *Статистичне опрацювання кількості абзаців у творах М. Яцкова та В. Стефаника*

Показник	**М. Яцків**	**В. Стефаник**
Кількість дослідів	42	55
Максимальне значення	203	158
Мінімальне значення	6	6
Середнє значення	38,48	36,05
Середнє квадратичне відхилення	42,44	30,24
Міра коливання середньої частоти	6,55	4,08
Стандартна похибка відхилення	6,63	4,11
Відносна похибка	0,33	0,22

Таблиця 10. *Довжини абзаців у творах М. Яцкова та В. Стефаника*

Показник	**М. Яцків**	**В. Стефаник**
Кількість різних значень	17	20
Середнє значення (кількість речень)	2,45	2,27

Показник	М. Яцків	В. Стефаник
Середнє квадратичне відхилення	2,22	2,24
Міра коливання середньої частоти	0,06	0,05
Стандартна похибка відхилення	0,06	0,05
Відносна похибка	0,04	0,04

Таблиця 11. *Статистичне опрацювання довжини речень у творах М. Яцкова та В. Стефаника*

Показник	М. Яцків	В. Стефаник
Кількість різних значень	52	50
Середнє значення (кількість словоформ)	10,73	11,19
Середнє квадратичне відхилення	7,78	6,95
Міра коливання середньої частоти	0,12	0,1
Стандартна похибка відхилення	0,12	0,1
Відносна похибка	0,02	0,02

Таблиця 12. *Статистичне опрацювання довжини словоформи в новелах М. Яцкова та В. Стефаника*

Показник	М. Яцків	В. Стефаник
Кількість значень	19	16
Середнє значення	4,75	4,31
Середнє квадратичне відхилення	2,64	2,36
Міра коливання середньої частоти	0,01	0,01
Стандартна похибка відхилення	0,01	0,01
Відносна похибка	0,01	0

Результати аналізу свідчать про зіставність показників у обох авторів. Брак даних про твори інших авторів не дає змоги зробити більш детальний аналіз. Зазначимо тільки, що у творах Михайла Яцкова порівняно з творами Василя Стефаника більша кількість абзаців та їхня довжина в реченнях. Менша довжина речення у словоформах, але більша їхня довжина у символах.

Отже, проведені дослідження поставили більше запитань, ніж дали відповідей. Найважливіша відповідь – квантитативні дослідження текстів українською мовою треба продовжувати, розширюючи кількість авторів, кількість творів, упроваджуючи нові методи та методики. Звичайно, можна поставити запитання: навіщо? Адже для більшости розвинених мов такі дослідження проведені й відомі. Чи не винаходимо ми наново велосипед? На глибоке моє переконання, дослідження проводити належить обов'язково. Кожна мова має феноменологічні властивості, й не можна результати, здобуті під час вивчення однієї мови повністю переносити на інші. Звичайно, спільні закономірності є, але диявол, зазвичай, прихований у деталях. А українська мова статистично обстежена надзвичайно мало. Work must go on.

Ihor KULCHYTSKYY

Particular Aspects of the Quantitative Studies of the Ukrainian Language

Ihor KULCHYTSKYY *– Ph.D., Assoc.Prof., Department of Applied Linguistics of Lviv Polytechnic National University. Interests: corpus lingistics, computational linguistics, lexicography, Ukrainian texts published in the Western Ukraine before 1939–1944.*

This article describes the various calculations of the number of linguistic phenomena; the construction of their stochastic models based on these quantitative data; the testing of hypotheses about these various phenomena using statistical methods; and implementation of the obtained results in various fields related to the use and learning of natural languages. Each linguistic unit is researched as a component of a language system. Studies of earlier and newer collections of texts have made it possible to confirm both previously defined and new statistical parameters and patterns of textual structuring as well as the structural features of various languages. The article's particular focus is on text attribution (authorship). Provided are the results of statistical studies of literary works by Vasyl Stefanyk, Mykhailo Iatskiv, Lina Kostenko, and Petro Karman'skyi carried out in the Department of Applied Linguistics of Lviv Polytechnical National University. Calculated were the frequency of alphabetic symbols and their euphony, and the required minimum sample size vis-à-vis the entire volume of an author's texts is determined. It is posited that such samples must be formed by means of the random selection of textual fragments from all of an author's works. Our tests obtained a negative result regarding the hypothesis that the frequency of such symbols may indicate authorship. The number of paragraphs in Stefanyk's and Iatskiv's works, of the number of sentences in each of their paragraphs, of the number of word forms in the sentences, and the word forms' length were statistically established.

Keywords: linguistic statistics, Ukrainian language, Vasyl Stefanyk, Mykhailo Iatskiv, Lina Kostenko, Petro Karmans'kyi, frequency, quantitative research.

Bibliography

Batura, T. V. "Formal'nye metody opredeleniia avtorstva tekstov". *Vestnik Novosib. gos. un-ta. Seriia: Informatsionnye tekhnologii*,10, №4 (2012): 81–94.

Buk, Solomiia. "Linhvostatystychnyi opys "Ne spytavshy brodu" Ivana Franka". *Visnyk L'vivs'koho universytetu*, 55 (2011): 230–242.

– "Statystychni kharakterystyky leksyky osnovnykh funktsional'nykh styliv ukrains'koi movy: sproba porivniannia". *Leksykohrafichnyi biuleten'*, 13 (2006): 166–170.

– "Suchasni metody doslidzhennia movy pys'mennyka u slov'ianoznavstvi". *Problemy slov'ianoznavstva*, 61(2012): 86–95.

Iaglom, A. M., i I. M Iaglom. *Veroiatnost' i informatsiia*. Moskva: KomKniga, 2007.

Iatskiv, Mykhailo. *Vybrani tvory*. Kyiv: Dnipro, 1973.

Karmans'kyi, P. S. *Poezii*. Kyiv: Ukr. pys'mennyk, 1992.

Köhler, Reinhard, Gabriel Altmann, and Rajmund Piotrowski, eds. *Quantitative Linguistik : ein internationals Handbuch = Quantitative linquistics : an international handbook*. Berlin, New York: de Gruyter, 2005.

Kostenko, Lina. *Vybrane*. Kyiv: Dnipro, 1989.

Kul'chyts'kyi, Ihor. "Doslidzhennia dovzhyny rechennia ta slova u tvorakh Romana Ivanychuka". *Visnyk Natsional'noho universytetu "L'vivs'ka politekhnika": Informatsiini systemy ta merezhi,* 872 (2017): 139–149.

– "Entropiia odno- ta dvohram symvoliv v ukrainomovnykh tekstakh". *Naukovyi visnyk Volyns'koho natsional'noho universytetu imeni L. Ukrainky,* 6 (2016): 183–189.

Mova.info "Linhvistychnyi portal". Accessed February 24, 2019. *http://www.mova.info/*

Perebyinis, V. I. "Shcho daie statystyka linhvistam?" *Visnyk Kyivs'koho linhvistychnoho universytetu* VI, № 2 (2003): 27–29.

Perebyinis, Valentyna, red. *Statystychni parametry styliv*. Kyiv: Naukova dumka, 1967.

Ruzhevych, N. A. *Matematychna statystyka*. L'viv: Vydavnytstvo Natsional'noho un-tu L'vivs'ka politekhnika, 2001.

Shyrokov, V. A. *Informatsiina teoriia leksykohrafichnykh system*. Kyiv: Dovira, 1998.

Stefanyk, Vasyl'. "Mezha". *Literaturno-naukovyi visnyk,* 92, № 2 (1927): 97–98.

– "Portret". U kn. *Tvory*, 94–95. Kharkiv: DVU, 1929.

– "Shkil'nyk". *Ridna shkola* (1932): 2–4.

– *Tvory*. L'viv: Vydavnycha Spilka "Dilo", 1933.

The International Quantitative Linguistics Association. Accessed February 24, 2019. http://www.iqla.org/

Zade, Lotfi. *Poniatie lingvisticheskoi peremennoi i ego primenenie k priniatiiu priblizhennykh reshenii.* Moskva: Mir, 1976.

Zvegintsev, V. A. *Ocherki po obshchemu iazykoznaniiu*. Moskva: MGU, 1962.