КОМП'ЮТЕРНІ НАУКИ

ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2022.01.037
Article received 06.05.2022 p.
Article accepted 19.05.20221 p.
UDK 004.4-22:003.08

Correspondence author
V. M. Teslyuk
vasyl.m.teslyuk@lpnu.ua

В. М. Теслюк, І. Я. Казимира, Ю. М. Кордіяка, І. Р. Рибак
Національний університет "Львівська політехніка", м. Львів, Україна

МОДЕЛІ ТА ЗАСОБИ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СТАТИСТИЧНОГО ПРОФІЛЮ УКРАЇНОМОВНИХ ТЕКСТІВ

У роботі вирішується актуальне завдання із вдосконалення професійного програмного забезпечення для статистичного аналізу тексту відповідно до потреб фахівців. Проаналізовано особливості і перспективи статистичних досліджень у мовознавстві та розроблено інформаційну технологію (ІТ) визначення статистичного профілю україномовних текстів. Проведено комплексну роботу над моделюванням програмної системи, яку представлено у відповідних схемах і діаграмах, що цілісно відображають функціонування та призначення розробленого продукту. Розглядаються математичні та системні основи статистичного аналізу для автоматизації професійного опрацювання текстів українською мовою, в контексті впровадження пропонованої інформаційної технології. Побудовано структурну схему проектного рішення та визначено головні вимоги до апаратного забезпечення. Розроблено компоненти інформаційної технології та запропоновано структуру програмної системи, які ґрунтуються на модульному принципі. Розроблено математичне забезпечення ІТ, яке базується на методах прикладної статистики та дає змогу визначити основні характеристики (статистичний профіль) досліджуваних україномовних текстів. Окрім цього, розроблено алгоритмічне та програмне забезпечення ІТ, для реалізації якого використано Python. Наведено результати дослідження україномовних текстів та їх статистичні профілі, продемонстровано, що розроблена інформаційна технологія забезпечує опрацювання україномовних текстів з високим рівнем автоматизації. Отримані результати можна розглядати як внесок у розвиток наукових досліджень у лінгвістиці, завдяки якому створюються умови для вивчення авторських текстів різного стилю та ефективного використання професійних навичок та знань широким колом користувачів.

Ключові слова: опрацювання даних; статистичний аналіз; лінгвістика тексту; інформаційна система; автоматизація.

Вступ/Introduction

Важливим аспектом в історії писемності, який вивчають з різних наукових позицій, є перехід на електронне письмо, що зберігається у пам'яті машин. Інформаційно-технологічні досягнення людства вливаються у процеси подальшого становлення та розвитку писемності, тому актуальним є постійне оновлення відомих і загальноприйнятих у лінгвістиці підходів і методів дослідження. Використання тільки людських здібностей і можливостей при статистичному опрацюванні текстової інформації, на сьогодні, є неефективним та потребує значних витрат часу в процесі виконання громіздких обчислень.

Проблематика комплексного вивчення текстів є основою лінгвістичного наукового середовища, на якій зосереджується розвиток мовознавства, особливо національної мови як засобу самоідентифікації. Наявні системи дослідження тексту, зазвичай, використовують тільки для англійської мови. При детальному вивченні напряму дослідження встановлено, що для дослідження україномовного тексту не існує готових програмних рішень з високим рівнем автоматизації, які б відповідали потребам та запитам користувачів. Для дослідження текстів використовують тільки окремі застарілі програми, тому актуальним завданням сьогодення є реалізація комплексного рішення для впровадження інформаційних технологій та систем автоматизації опрацювання статистичних даних для розвитку досліджень у лінгвістиці.

Об'єкт дослідження – автоматизоване визначення статистичного профілю україномовних текстів.

Предмет дослідження – моделі та засоби автоматизованого визначення статистичного профілю україномовних текстів.

Мета роботи – розроблення інформаційної технології для автоматизації визначення статистичного профілю україномовних текстів.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- провести аналіз наявних програмних рішень для автоматизації дослідження україномовного тексту;

- побудувати структурну схему інформаційної технології автоматизації визначення статистичного профілю україномовних текстів;
- розробити моделі для реалізації автоматичного опрацювання статистичної текстової інформації;
- виконати тестування програмного засобу на відповідність поставленим вимогам.

Аналіз останніх досліджень та публікацій. Для класифікації та опрацювання текстів у лінгвістиці використовують чотири основні групи методів, які працюють на морфологічному, статистичному, синтаксичному, семантичному рівнях, відповідно [5].

Статистичний аналіз текстової інформації дослідники і науковці, реалізовують на основі математично-статистичного та асоціативно-статистичного підходів. Останній ґрунтується на природному накопиченні асоціацій між образами та закріпленні рефлексів шляхом повторень, що не потребує попередньої підготовки із залученням висококваліфікованих експертів [1]. Стимулом для активного використання математичних підходів у вивчені мови є перспективи машинного перекладу, оскільки при становленні процесу опрацювання текстових одиниць одержано різноманітні кількісні оцінки окремих ознак мови, які згодом виявилися корисними не тільки для створення математичних моделей, а й для лінгвістичної теорії [4]. Можливість використання кількісних статистичних оцінок у мовознавстві ґрунтується на особливостях будови мови та мовлення.

Мова – це система, яка складається з дискретних одиниць, що мають кількісні характеристики та ймовірнісний характер. Відповідно до [7] визначено чинники, що дають змогу застосовувати кількісні методи при дослідженні мовних і мовленнєвих даних: дискретність одиниць, масовість мовних одиниць, повторюваність у висловлюваннях, можливість вибору певного елемента з набору однорідних, тощо. Також потрібно врахувати те, що на формування мовлення впливають закономірності будови та використання одиниць мови, закони їхньої сполучуваності, специфіка жанру, теми висловлювання, смаки автора, його психофізіологічний стан, проте, якщо сукупність цих чинників відносно постійна, то будова мовлення набуває характерних рис, тому розкриття закономірностей функціонування одиниць мови у мовленні, а також встановлення закономірностей будови тексту можна вивчати кількісними методами, що є основним завданням статистичної лінгвістики [11], [23].

У мовознавстві важливими об'єктами дослідження є тексти і корпуси текстів [10], [12]. Якщо текст є сукупністю зв'язаних і послідовних символів, то корпус являє собою великий обсяг текстів, які відібрані за певними правилами для статистичного аналізу і перевірки гіпотез, підтвердження лінгвістичних правил та дослідження особливостей мови. Аналіз корпусів для української мови переважно виконується в закритому доступі, проте має чітко визначені характеристики [2]. Оскільки всі рівні мовної системи підпорядковуються дії статистичних законів, то опрацюванню можуть підлягати одиниці будь-якого рівня – фонеми або звуки, літери, сполуки звуків, фонем або літер, склади, морфеми, слова, словосполучення, синтаксичні конструкції тощо. Підготовка тексту до аналізу здійснюється за визначеним алгоритмом, залежно від обраного підходу для реалізації статистичного аналізу [17].

Розрізняють описову статистику та статистичні висновки. До описової статистики можна віднести взаємозв'язки між змінними. Вона дає можливість описати досліджуваний об'єкт, отримавши нову інформацію і оцінити її. До статистичних висновків належить статистичне оцінювання – перевірка на точність, ефективність і надійність вибірки, а також виявлення похибки при дослідженні [23]. Ще однією можливістю статистичних висновків є перевірка статистичних гіпотез [9].

Традиційні методи лінгвістики для аналізу тексту дають змогу виконати поставлені завдання, проте підвищити достовірність отриманих результатів можна тільки з використанням машинного аналізу на основі обчислювального інтелекту. Враховуючи основні завдання інтелектуального аналізу [6], для початку роботи потрібно представити текст у форматі, зручному для програмних засобів. Важливо, що при цьому потрібно проаналізувати значну кількість систем та бібліотек, котрі можуть працювати з текстами на поширених мовах світу, тому можливості вітчизняної комп'ютерної лінгвістики істотно обмежені. З огляду на це, постає необхідність розроблення та реалізації математичного та програмного забезпечення для опрацювання та дослідження саме україномовних текстів. У цьому напрямі активно проводять дослідження для вирішення окремих питань, що описані у низці робіт [3], [14], [20], [22], а також українських науковців цікавлять можливості статистичного опрацювання іншомовних текстів [8].

Зазначивши позитивні аспекти згаданих вище досліджень, потрібно підкреслити необхідність їх подальшого структурного розвитку, оскільки аналіз наявного стану інформаційного-технічного забезпечення для статистичного дослідження україномовних текстів доводить, що основні наукові результати зосереджені на теоретичному рівні без подальшої практичної реалізації. Наявні інформаційні технології у лінгвістиці україномовних текстів істотно застарілі, тому їхнє оновлення є логічним продовженням концепції досліджень інших науковців, що відображено у меті роботи.

Проведений аналіз дає змогу стверджувати, що більшість доступних програм для статистичного аналізу україномовного тексту мають істотні обмеження у виборі мови та застарілий інтерфейс, тому їхні критерії функціональної спроможності не відповідають рівню вимог для активного професійного використання. Вивчення аналогів розробленого програмного рішення доводить, що питання покращення дизайну подібних систем для зручності і адаптивності у роботі з тестами української мови є актуальним завданням і потребує істотного вдосконалення.

Результати дослідження та їх обговорення / Research results and their discussion

Розроблення інформаційної технології автоматизованого визначення статистичного профілю україномовних текстів. Будь-яка автоматизована інформаційна система характеризується наявністю технології перетворення вихідних даних у інформацію нової якості, тому структурування інформаційної технології пов'язане з розробленням системи класифікації та кодування, організацією збирання та передавання інформації, та пошуком різних методів доступу до даних. Зокрема, розроблена структура інформаційної технології

для визначення статистичного профілю україномовних текстів зображена на рис. 1.

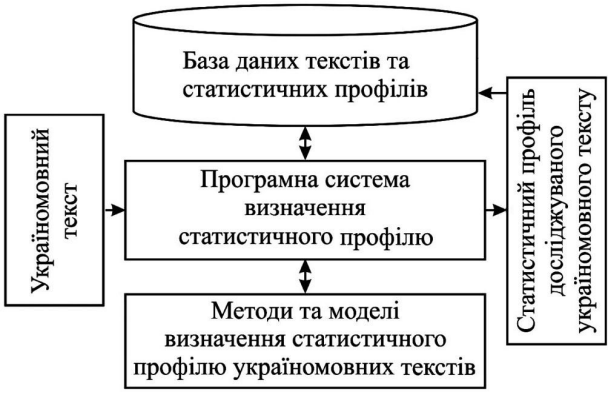

Рис. 1. Структура інформаційної технології визначення статистичного профілю україномовних текстів / The structure of information technology for determining the statistical profile of Ukrainian-language texts

Побудована інформаційна технологія передбачає опрацювання з використанням статистичних методів та програмних засобів набору досліджуваних текстів певного автора. Внаслідок цього, засоби інформаційної технології визначення статистичного профілю україномовних текстів формують документ із статистичними характеристиками досліджуваного тексту, притаманними автору тексту, тобто такими, які складають його статистичний профіль.

Математичне та алгоритмічне забезпечення інформаційної технології автоматизованого визначення статистичного профілю україномовних текстів. У процесі реалізації засобів ІТ автоматизованого визначення статистичного профілю україномовних текстів використано мову Python. Відповідно, реалізація ряду математичних моделей для опрацювання текстових одиниць виконана з використанням стандартних функцій, бібліотек та засобів Python. Зокрема, використані інструменти Python для роботи з різними мережевими протоколами, наприклад, модулі для роботи з XML та інші.

Процес визначення статистичного профілю передбачає виконання двох основних етапів, а саме: перший етап полягає в підготовці тексту, а другий – безпосереднє визначення статистичних параметрів. Математичну модель визначення статистичного профілю можна зобразити з допомогою такого кортежа:

$$П_{профіль\ тексту} = < T_i, M_{опр}, M_{стат}, П_i >$$

де: T_i – множина досліджуваних текстів; $M_{опр}$ – множина моделей попереднього опрацювання текстових даних; $M_{стат}$ – моделі та методи статистичного опрацювання досліджуваних текстів; $П_i$ – профіль i-го досліджуваного тексту.

Отже, опрацювання текстової інформації в ІТ автоматизованого визначення статистичного профілю україномовних текстів передбачає попередню підготовку досліджуваного тексту. Для таких задач використовуються моделі і методи пошуку, видалення символів (лінгвістичних об'єктів) та їх сортування. Для прикладу, у процесі роботи зі стоп-словами розроблено відповідні моделі. У пропонованій ІТ для автоматизованого визначення статистичного профілю україномовних текстів існує можливість початково опрацьовувати такі типові загальні стоп-слова:

- цифри: 1, 2, 3, 4, 5, 6, 7, 8, 9, 0 (один, два, три, чотири, п'ять, шість, сім, вісім, дев'ять, нуль);
- пунктуаційні знаки, що стоять окремо;
- букви розташовані окремо;
- займенники, дієприкметники, прийменники, вигуки, суфікси і поєднання букв: без, більш, б, був, була, були, було, бути (окрім фразеологічних зворотів, таких як "бути чи не бути"), вам, вас, адже, весь, вздовж, замість, поза, вниз, внизу, всередині, під, навколо, от, все, завжди, все, всіх, ви, де, да, навіть, для, до і т. д.;
- слова, які часто зустрічаються в обраній темі;
- нецензурна мова.

На наступних етапах опрацювання текстової інформації розроблено моделі та методи, які ґрунтуються на теорії прикладної статистики, вони детально описані в роботах [13], [16], [18], [21].

Алгоритм функціонування програмної системи передбачає такі основні кроки: завантаження файлу, запис стоп-слів, ідентифікація ключових позначень, поділу тексту на масиви, знаходження унікальних слів, статистичне опрацювання текстових одиниць, пошук та виведення інформації.

Засоби інформаційної технології автоматизованого визначення статистичного профілю україномовних текстів. Розроблена структура програмної системи визначення статистичного профілю зображена на рис. 2. Структурна схема системи містить основні модулі, а саме: модуль відкриття файлу; модуль налаштування вибірки; модуль попереднього аналізу досліджуваного тексту (групи текстів); модуль візуалізації результатів; модуль збереження даних.

Модуль збереження даних виконує доступ до бази даних, де безпосередньо зберігаються досліджувані тексти та профілі. Окрім цього, зв'язок між користувачами та програмною системою забезпечено з допомогою спеціального модуля "Інтерфейс". Необхідно звернути увагу, що дана структура базується на модульному принципі, що дає змогу швидкої модифікації системи.

Рис. 2. Графічне зображення структурної схеми програмної системи / Graphic representation of the structural diagram of the software system

Для автоматизації опрацювання текстових одиниць та профілів розроблено базу даних, структура якої зображена на рис. 3. Для збереження даних про досліджуваний текст (статистичний профіль) використовується описова структура даних, яку показано на рис. 4.

Як зазначено вище, програмну реалізацію засобів ІТ для попереднього опрацювання та визначення статистичного профілю текстів здійснено за допомогою мови Python, що є високорівневим інструментом програмування із численною кількістю бібліотек для проведення статистичного аналізу даних. Потрібно зауважити, що для програмного опрацювання текст необхідно перет-

ворити у зручну і зрозумілу форму для однозначної машинної ідентифікації. Для цього використано простий та ефективний підхід − рядковий метод split(), оскільки додаткові бібліотеки, які здійснюють символьну і статистичну обробку природної мови, а також містять графічні представлення і приклади даних, у стандартному режимі роботи не опрацьовують українські слова і символи. Для залучення додаткових сервісів потрібно виконувати попереднє оброблення і очищення тексту, що значно погіршує показник швидкодії етапу поділу тексту на окремі мовні одиниці для аналізу.

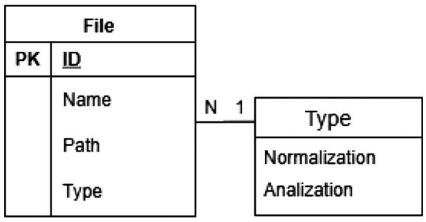

Рис. 3. Схема бази даних / Database scheme

Рис. 4. Приклад описової структури даних і профіль досліджуваного тексту / Example of descriptive data structure and profile of the researched text

На рис. 5 зображено діаграму компонент, яка відображає сукупність зв'язаних елементів програмного і апаратного забезпечення.

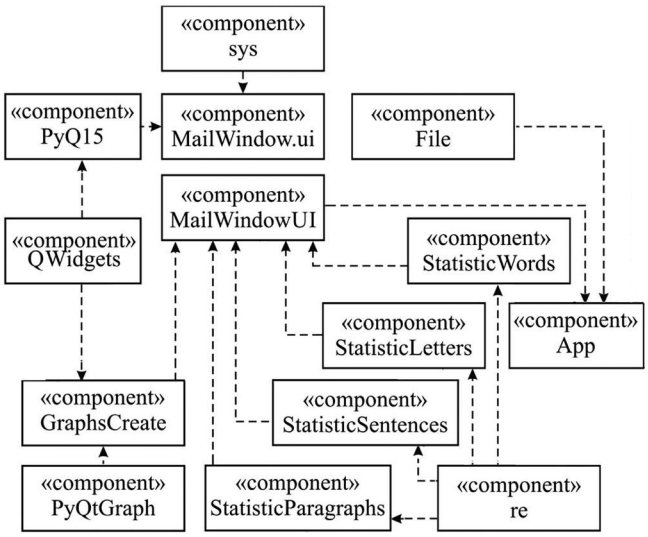

Рис. 5. Діаграма компонент IT / IT component diagram

Під час програмної розробки засобів IT використано низку технологій, враховуючи особливості запропонованого програмного забезпечення. Інтерфейс програми написаний за допомогою бібліотеки PyQt5, яка містить дизайнер графічного інтерфейсу користувача (Qt Designer). Програма pyuic генерує Python-код з файлів, створених в Qt Designer, що робить PyQt5 корисним інструментом для швидкого прототипування. Бібліотека Qt є однією з найпотужніших бібліотек графічного інтерфейсу користувача, тому PyQt5 має достатньо модулів та класів для розроблення інтерфейсу програми. Одним із найбільших класів даної бібліотеки є QWidgets, він є базовим для всіх об'єктів користувальницького інтерфейсу, оскільки містить усі таблиці, списки і інші засоби для візуалізації результатів програми. Графічна бібліотека та інтерфейс PyQtGraph побудовані на PyQt та numpy для використання в математичних, наукових, інженерних програмах. Незважаючи на те, що бібліотека написана повністю на Python, вона працює дуже швидко завдяки значному рівню numpy та фреймворку GraphicsView Qt для швидкої візуалізації.

Для зручної та ефективної роботи запропонованого програмного рішення розроблено окремі модулі: MainWindowUI, який відображає головне і дочірні вікна програми та імпортує всі наступні модулі, а також файл MainWindow.ui, в якому знаходиться код інтерфейсу програми, StatisticLetters, StatisticWords, StatisticSentences, StatisticParagraphs, котрі відповідають за статистичний аналіз букв, слів, речень та абзаців. Модуль re здійснює редагування та підготовку тексту до опрацювання, а GraphsCreate містить функції для графічного представлення отриманої інформації. File − збереження тексту, модуль sys − доступ до деяких змінних і функцій, які взаємодіють з інтерпретатором Python.

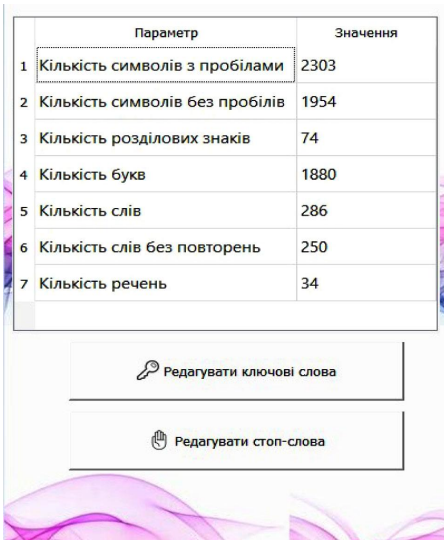

Рис. 6. Основні параметри статистичного аналізу тексту / The main indicators of the text statistical analysis

Для перевірки результатів роботи програми визначено та описано головні вимоги до апаратного забезпечення, доведено повне та правильне виконання задуманого функціоналу та коректність роботи реалізованого застосунку. Важливо, що інтерфейс програми є простим та зрозумілим у користуванні, а графічна інтерпретація результатів відповідає математичним закономірностям. Зокрема, на рис. 6 зображено результати дослідження статистичного аналізу досліджуваного тексту. Приклад меню для опрацювання стоп-слів наведено на рис. 7. Зібравши максимально повний перелік стоп-слів різного характеру, можна ще більше скоротити час на вичитку унікальних слів. Пов'язано це з тим, що після додавання списку унікальних слів таблиця автоматично перевіряє їх на відповідність стоп-словами. Якщо відповідність знайдено, ставиться відмітка. Ви можете видаляти, змінювати і додавати свої списки стоп-слів. Але доведеться розібратися в тому, як все працює.

У вкладці "Символи системи" представлені досліджувані мовні одиниці та частота їх появи в тексті (рис. 8). Саме ці знаки складають основу мови. Слова і є знаками-символами. Смислом є інформація, яку несе знак про предмет. Тут необхідно зауважити, що, говорячи про смисл знака, маємо на увазі інформацію про предмет, завдяки якій ми однозначно виділяємо предмет і відрізняємо його від інших предметів. Тобто не будь-яка інформація про предмет може відігравати роль смислу. Таку інформацію називають прямим смислом.

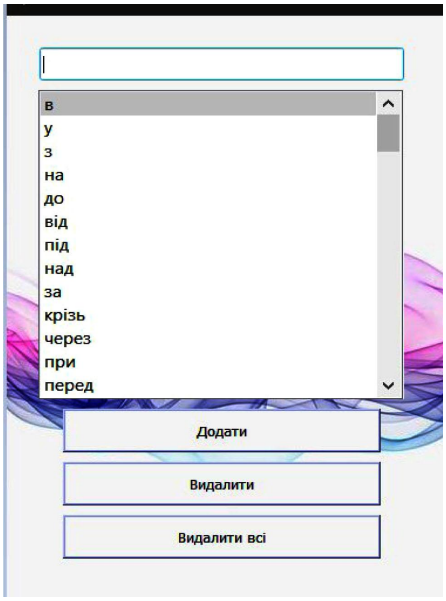

Рис. 7. Вікно для редагування стоп-слів та ключових слів / Window for stop words and keywords editing

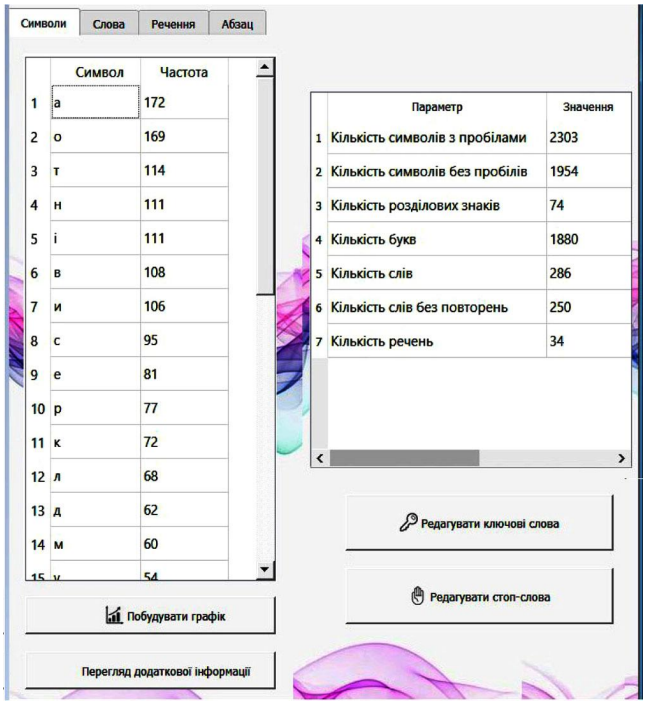

Рис. 8. Виведення результатів статистичного аналізу тексту / Output of the results of the text statistical analysis

На кожній вкладці інтерфейсу системи знаходяться опції для побудови відповідних графіків, які відображають статистичні результати аналізу. Графіки мають додаткові функції редагування, налаштування (рис. 9) та збереження отриманої інформації. Також є опції для відображення додаткової інформації, яка відкривається у новому вікні. Після перегляду результату, користувач зможе зберегти його в файл (текст *.txt або електронні таблиці *.xls), окрім цього, при виході система автоматично помістить файл з отриманими результатами в базу даних.

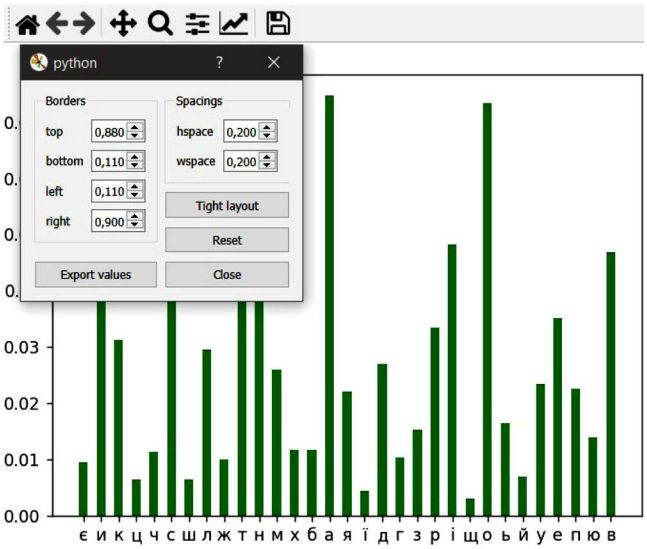

Рис. 9. Можливості налаштування графічного представлення результатів статистичного аналізу / Setting options for graphical representation of the statistical analysis results

Обговорення отриманих результатів. Проведено тестові випробування розробленої системи для побудови статистичного профілю україномовних текстів, що підтверджують її функціональну відповідність поставленим вимогам. Розглядаються математичні та системні основи статистичного аналізу для автоматизації професійного опрацювання текстів українською мовою, в контексті впровадження пропонованої інформаційної технології. Побудовано структурну схему проектного рішення та визначено головні вимоги до апаратного забезпечення.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів – розроблено модель автоматизованого визначення статистичного профілю текстів української мови, яка забезпечує можливість комплексного дослідження корпусу текстів.

Практична значущість отриманих результатів – розроблено та реалізовано програмні засоби інформаційної технології автоматизації визначення статистичного профілю україномовних текстів, структурна схема ІТ та отримані результати дослідження.

Висновки / Conclusions

Розроблено структуру інформаційної технології та засобів інформаційної технології для автоматизованого визначення статистичного профілю україномовних текстів, яка базується на модульному принципі. Розроблені засоби ІТ забезпечують автоматизацію докорпусної підготовки та статистичного опрацювання з представленням результатів у графічному і табличному вигляді, що істотно полегшує роботу фахівців при вирішенні окремих завдань.

Побудовано модель визначення статистичного профілю україномовних текстів. Вдосконалено методи та моделі для визначення статистичного профілю україно-

мовних текстів, які ґрунтуються на теорії прикладної статистики, що дають змогу в автоматизованому режимі визначити характеристики профілів україномовних текстів.

Розроблено інформаційне та програмне забезпечення для автоматизованого визначення статистичного профілю україномовних текстів, для реалізації якого використано Python.

Проведено тестування розроблених засобів інформаційної технології, отримані результати підтверджують правильність та коректність їх роботи з україномовними текстами.

References

[1] Bisikalo, O. V., & Kravchuk, I. A. (2010, November). Analysis of the morphological structure of the word based on the associative-statistical approach. *Journal of Vinnytsia Polytechnic Institute, 4*, 134–136. Retrieved from: www.visnyk.vntu.edu.ua/index.php/visnyk/article/view/1495

[2] Buk, S. N., & Rovenchak, A. A. (2004). Rank-Frequency Analysis for Functional Style Corpora of Ukrainian. *Journal of Quantitative Linguistics, 11*(3), 161–71. https://doi.org/10.1080/0929617042000314912

[3] Grabar, N., & Thierry, H. (2017, April). Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. *Computational linguistics and intelligent systems* (*COLINS 2017*)*: proceedings of the 1st International conference, National Technical University "KhPI",* 10–19. Retrieved from: http://ena.lp.edu.ua:8080/handle/ntb/39454

[4] Grodniewicz, J. P. (2021). The process of linguistic understanding. *Synthese, 198*, 11463–11481. https://doi.org/10.1007/s11229-020-02807-9

[5] Hlushchenko, V. A. (2010). Linguistic method and its structure. *Linguistics, 6*, 32–44. Retrieved from: http://nbuv.gov.ua/UJRN/MoZn_2010_6_5

[6] Hlybovets, A. M., & Tochytsky, V. V. (2017). Algorithm of tokenization and steaming for texts in Ukrainian. *NaUKMA Research Papers Computer Science, 198*, 4–8. Retrieved from: http://nbuv.gov.ua/UJRN/NaUKMAkn_2017_198_4

[7] Hoherchak, H., Darchuk, N., & Kryvyi, S. (2021). Representation, Analysis, and Extraction of Knowledge from Unstructured Natural Language Texts. *Cybern Syst Anal, 57*, 481–500. https://doi.org/10.1007/s10559-021-00373-7

[8] Khomytska, I. Y., Teslyuk, V. M., Bazylevych, I. B., & Beregovskyi, V. V. (2020). The statistical models and software for authorial style differentiation in english prose. *Scientific Bulletin of UNFU, 30*(5), 135–139. https://doi.org/10.36930/40300522

[9] Lawson, A. E., Oehrtman, M., & Jensen, J. (2008) Connecting Science and Mathematics: The Nature of Scientific and Statistical Hypothesis Testing. *Int J of Sci and Math Educ, 6*, 405–416. https://doi.org/10.1007/s10763-007-9108-5

[10] Levchenko, O., & Dilai, M. (2021). A Method of Automated Corpus-Based Identification of Metaphors for Compiling a Dictionary of Metaphors: A Case Study of the Emotion Conceptual Domain. *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies* (*CSIT*), 52–55. https://doi.org/10.1109/CSIT52700.2021.9648667

[11] Levchenko, O., Holtvian, V., & Dilai, M. (2021). Statistical profiles of Ukrainian prose fiction: Gender aspect. *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies* (*CSIT*), 97–100. https://doi.org/10.1109/CSIT52700.2021.9648668

[12] Levchenko, O., Tyshchenko, O., & Dilai, M. (2021). Automated identification of metaphors in annotated corpus (Based on substance terms). *CEUR Workshop Proceedings, 2870*(3), 16–31. Retrieved from: http://ceur-ws.org/Vol-2870/paper3.pdf

[13] Lupenko, S. A., Khomiv, B. A., & Sverstyuk, A. S. (2011) Comparative analysis of mathematical models, methods and methods for evaluating opinions in text data from Internet resources. *Bulletin of Khmelnytsky National University. 6,* 7–16. Retrieved from:http://ceur-ws.org/Vol-2870/paper3.pdf http://journals.khnu.km.ua/vestnik/zmisthtm/2011-6-t.htm

[14] Lytvyn, V., Vysotska, V., Uhryn, D., Hrendus, M., & Naum, O. (2018). Analysis of statistical methods for stable combinations determination of keywords identification. *Eastern-European Journal of Enterprise Technologies, 2* (2 (92)), 23–37. https://doi.org/10.15587/1729-4061.2018.126009

[15] Nikonenko, A. O. (2012). Review of computer-linguistic methods of processing natural language texts. *Artificial Intelligence, 4*, 235–244. Retrieved from: http://dspace.nbuv.gov.ua/handle/123456789/57737

[16] Ostapova, I.V., Shirokov, V.A., Luchik, A. A., & Yablochkov, N. M. The study of the functioning of word equivalents in the text on the material of the Ukrainian National Linguistic Corpus. *Speech Technology,* (*1-2*)*,* 114–120.

[17] Parshak, K. D. (2014). Text as an object of linguistic research. Scientific journal of M. P. Dragomanov National Pedagogical University. Series 10: Problems of grammar and lexicology of the Ukrainian language, 11, 196–199. Retrieved from: http://nbuv.gov.ua/UJRN/Nchnpu_10_2014_11_46

[18] Perebyinis, V. S., (1967) *Statistical style settings.* Kyiv: Naukova Dumka.

[19] Romaniuk, S. (2015). Application of statistical methods in linguistic research. *Scientific Proceedings of Ostroh Academy National University: Philology Series, 54*, 134–137. Retrieved from: http://eprints.oa.edu.ua/id/eprint/4185

[20] Rovenchak, A., & Buk, S. (2011). Application of a quantum ensemble model to linguistic analysis. *Physica A: Statistical Mechanics and its Applications, 390*(7), 1326–1331. https://doi.org/10.1016/j.physa.2010.12.009

[21] Shyrokov, V., Ostapova, I., &Yakymenko, K. (2014) Indexing the etymological lexicographic systems Cognitives Studies. *Warsaw : SOW Publishing House*, 13–23. https://doi.org/10.11649/cs.2014.001

[22] Tkachenko, O., & Humeniuk, M. (2020). Aspects of visualization of statistical and scientific data. *Digital platform: information technologies in the socio-cultural sphere, 3*(2), 134–147. https://doi.org/10.31866/2617-796x.3.2.2020.220584

[23] Zaiats, V. M., & Zaiats, M. M. (2010). Methods of comparing statistical characteristics in the formation of samples in linguistics. *Journal of Lviv Polytechnic National University "Information Systems and Networks", 673*, 296–305. Retrieved from http://ena.lp.edu.ua:8080/bitstream/ntb/6753/1/33.pdf

V. M. Teslyuk, I. Ya. Kazymyra, Yu. M. Kordiiaka, I. R. Rybak

Lviv Polytechnic National University, Lviv, Ukraine

MODELS AND TOOLS FOR AUTOMATED DETERMINING THE STATISTICAL PROFILE OF UKRAINIAN-LANGUAGE TEXTS

The paper deals with the urgent issue of improving the professional software for text statistical analysis in accordance with the needs of specialists. Peculiarities and prospects of statistical research in linguistics are analyzed and information technology (IT) for determining the statistical profile of Ukrainian-language texts is developed. Complex work on modelling the software

system was carried out, it was presented in the corresponding schemes and diagrams, which integrally reflect the functioning and purpose of the developed product. Mathematical and system bases of statistical analysis aimed at automation of professional processing of Ukrainian-language texts, in the context of introducing the offered information technology are considered. The structural scheme of the project decision is constructed and the main requirements for hardware are defined. The components of information technology are developed, and the software system structure is proposed, which is based on the modular principle. Mathematical support for IT has been developed, it is based on the methods of applied statistics and allows determining the main characteristics (statistical profile) of the studied Ukrainian-language texts. In addition, the algorithms and software for IT have been developed using Python. The results of research on Ukrainian-language texts and their statistical profiles are given, it is shown that the developed information technology provides processing of Ukrainian-language texts with a high level of automation. The obtained results can be considered as a contribution to the development of scientific research in linguistics, which creates conditions for the study of authors texts of different styles and the effective use of professional skills and knowledge by a wide range of users. The scientific novelty of the work is that a model of automated determination of the statistical profile of Ukrainian language texts has been developed, which provides an opportunity for a comprehensive study of the corpus of Ukrainian-language texts. The obtained results are also of practical significance, as the structural scheme of IT has been developed, software tools of information technology for automation of the determining the statistical profile of Ukrainian-language texts have been implemented, and the results of text investigation have been analyzed.

Keywords: data processing; statistical analysis; linguistics of the text; information system; automation.

Інформація про авторів:

Теслюк Василь Миколайович, д-р техн. наук, професор, завідувач кафедри автоматизованих систем управління.
Email: vasyl.m.teslyuk@lpnu.ua; https://orcid.org/0000-0002-5974-9310

Казимира Ірина Ярославівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління.
Email: iryna.y.kazymyra@lpnu.ua; https://orcid.org/0000-0003-1597-5647

Кордіяка Юлія Миронівна, канд. техн. наук, асистент, кафедра автоматизованих систем управління.
Email: yuliia.m.kordiiaka@lpnu.ua; https://orcid.org/0000-0002-5391-0556

Рибак Ірина, студентка магістратури, кафедра автоматизованих систем управління. **Email:** iryna.rybak.kn.2017@lpnu.ua.

Цитування за ДСТУ: Теслюк В. М., Казимира І. Я., Кордіяка Ю. М., Рибак І. Р. Моделі та засоби автоматизованого визначення статистичного профілю україномовних текстів. *Український журнал інформаційних технологій.* 2022, т. 4, № 1. С. 37–43.

Citation APA: Teslyuk, V. M., Kazymyra, I. Ya., Kordiiaka, Yu. M., & Rybak, I. R. (2022). Models and tools for automated determining the statistical profile of ukrainian-language texts. *Ukrainian Journal of Information Technology, 4*(1), 37–43. https://doi.org/10.23939/ujit2022.01.037