МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

АВТОМАТИЗАЦІЯ ЛІНГВОСТАТИСТИЧНОГО ЕКСПЕРИМЕНТУ: СТИЛЕМЕТРІЯ ТА ЛІНГВОДИДАКТИКА

Кваліфікаційна робота
студентки IV курсу
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»,
спеціальності 035 «Філологія»,
спеціалізації 035.10 «Прикладна лінгвістика»,
галузі знань 03 «Гуманітарні науки»
Уляни САСЮК

Науковий керівник:
к.філол.н., доц. кафедри
української мови
та прикладної лінгвістики
Оксана ЗУБАНЬ

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від 28.05.2025 № 15
Завідувач кафедри _____ **Сергій РІЗНИК**

КИЇВ – 2025

АНОТАЦІЯ

У кваліфікаційній роботі бакалавра досліджено теоретичні засади та практичне застосування квантитативної лінгвістики для аналізу функційних стилів української мови. Для чотирьох стилів (художній, науковий, офіційно-діловий, медійний) на основі корпусів по 50 000 словоформ було автоматично укладено частотні словники словоформ та частин мови. Усього було створено вісім частотних словників з дотриманням усіх необхідних вимог укладання вибірок та власне частотних словників. Частотні словники частин мови стали основою для стилеметричного дослідження, в межах якого було обчислено ключові статистичні характеристики вибірок кожного стилю. За отриманими результатами можна зробити висновки про значення кожної метрики для розрізнення стильової належності тексту, але саме зіставлення результатів за критеріями хі-квадрат і Стьюдента підтвердило суттєві відмінності між стилями. Одержані бази даних дозволяють детально охарактеризувати ці стилі лише за унікальними значеннями статистичних показників. Практичним результатом роботи стало створення інтерактивного застосунку для автоматичного генерування задач із квантитативної лінгвістики. Застосунок має навчально-методичне призначення та генерує задачі, які входять до курсу “Квантитативна лінгвістика”. Робота сприяє розвитку української стилеметрії та поширенню кількісних методів у мовознавстві. Надаються бази даних, які описують стан функційних стилів сучасної української мови, а також створено практичне втілення проведеного дослідження у вигляді інтерактивного застосунку. Отримані результати можуть бути використані в лінгводидактиці, стилістиці, комп’ютерній лінгвістиці та методиці викладання.

Ключові слова: квантитативна лінгвістика, статистичний метод, частотний словник, статистичний параметр, база даних, автоматичне генерування задач.

ANNOTATION

In this bachelor's qualification thesis, the theoretical foundations and practical applications of quantitative linguistics for the analysis of functional styles of the Ukrainian language were examined. For four styles (writing style, scientific writing, officialese, and news), word lists of word forms and parts of speech were automatically compiled based on corpora consisting of 50,000 word forms each. In total, eight word lists were created following all requirements for sampling and word list compilation. The word lists of parts of speech served as the basis for a stylometric study, within which the key statistical characteristics of the samples for each style were calculated. The obtained results allow for conclusions regarding the importance of each metric in distinguishing the stylistic affiliation of texts; however, it is the comparison of the results using the Xi-square and Student's t-test criteria that confirmed the significant differences between the styles. The resulting databases make it possible to characterize these styles in detail based on the unique values of statistical indicators. A practical outcome of the thesis is the development of an interactive application for the automatic generation of exercises in quantitative linguistics. The application is intended for educational and methodological use and generates tasks included in the course “Quantitative Linguistics.” This work contributes to the development of Ukrainian stylometry and promotes the use of quantitative methods in linguistic research. The thesis provides databases that describe the current state of the functional styles of modern Ukrainian language and implements the conducted research in the form of an interactive digital tool. The obtained results can be applied in language didactics, stylistics, computational linguistics, and language teaching methodology.

Key words: quantitative linguistics, statistical method, word list, statistical parameter, database, automatic problem generation.

Зміст

ВСТУП

Актуальність: квантитативні, зокрема статистичні методи, стали особливо актуальними у XXI столітті, адже вони дають можливість здійснити унікальні, а головне, підтверджені конкретними даними відкриття. З поширенням використання інформаційних технологій у лінгвістиці статистичні дослідження набули нового значення, активно розпочалась автоматизація різноманітних дослідницьких процесів, яка і полегшує роботу спеціалістів, і забезпечує точність та повноту висновків, і дозволяє працювати зі значно більшими масивами даних та досліджувати ті аспекти мови, які неможливо було проаналізувати глибинно до комп'ютеризації науки. Але для української мови фундамент таких досліджень лише формується. Так, багато дослідників та дослідниць активно впроваджують квантитативні методи у вивчення української мови. Це С. Бук, яка застосовує математичні методи у дослідженні текстів І. Франка [Бук 2001], Є. Карпіловська, яка ініціювала створення комп'ютерного морфемно-словотвірного фонду української мови [Клименко 2014], та Н. Дарчук, яка брала участь у створенні численних частотних словників, а також лінгвістичного порталу mova.info [mova.info 2025]. Проте необхідно і далі розвивати цю галузь науки в Україні для набуття нових, глибших та якісніших знань про рідну мову. Для розвитку науки завжди потрібні кваліфіковані спеціалісти, які мають не лише глибоке теоретичне розуміння мови, а й практичні навички, необхідні для проведення точних та вичерпних досліджень. Таку освіту сьогодні здобувають одиниці, тож варто ввести курси з квантитативної лінгвістики і до філологічних освітніх програм, аби більше студентів долучалися до поповнення наукової бази досліджень української мови.

У межах цієї роботи було проведено одне з таких досліджень, а саме створено частотні словники для чотирьох функційних стилів української мови, а також обчислено та порівняно їхні статистичні характеристики. Одержані результати та висновки можуть стати основою для подальших лінгвістичних досліджень у напрямку стилеметрії або лінгвостатистики, адже містять всю необхідну теоретичну та методологічну базу квантитативної лінгвістики. Також

ця робота дає приклад практичного застосування цих напрацювань – було створено продукт, який підтримуватиме вивчення квантитативної лінгвістики у нових студентів, які беруться за вивчення цієї галузі.

Мета: розробити систему автоматичного статистичного аналізу українськомовних текстів дослідницького та дидактичного спрямування.

Завдання:

1) дослідити історію та актуальність квантитативних методів у лінгвістиці;
2) створити програму автоматичного укладання частотних словників слів та частин мови на текстовому матеріалі чотирьох стилів української мови;
3) здійснити стилеметричне дослідження на базі укладених частотних словників;
4) створити інтерактивний дидактичний застосунок для генерування задач із курсу “Квантитативна лінгвістика”.

Об’єкт дослідження: тексти української мови.

Предмет дослідження: статистична структура українськомовних текстів; моделі статистичних обчислень базових статистичних характеристик та методів.

Методи дослідження: методи кількісного та статистичного аналізу, методика укладання частотних словників.

Мови програмування та інструменти візуалізації даних, використані у дослідженні: Python [Python 2025], PHP [PHP 2025], JavaScript [JavaScript 2025], HTML [html 2025] та CSS [CSS 2025], табличний процесор Microsoft Excel [Excel 2025], графічний інтерфейс DB Browser for SQLite [DB Browser].

Матеріал дослідження: тексти таких функційних стилів української мови: медійний – обсяг тексту – 50000 слововживань (див. Додаток 1), офіційно-діловий – обсяг тексту – 50000 слововживань (див. Додаток 1), художній – обсяг тексту – 50000 слововживань (див. Додаток 1), науковий – обсяг тексту – 50000 слововживань (див. Додаток 1). Загальний обсяг текстового матеріалу – 200000 слововживань.

Практичне та теоретичне значення роботи: теоретичне значення роботи – отримано нові стилеметричні дані для характеристики медійного, художнього,

офіційно-ділового і наукового стилів української мови; аргументовано необхідність застосування квантитативних методів у лінгвістичних дослідженнях. Практичне значення роботи – укладено частотні словники на основі чотирьох функційних стилів, які можуть бути використані в різноманітних лінгвостатистичних дослідженнях; проведено стилеметричне дослідження, результати якого можуть бути використані для автоматичного визначення стилю тексту; розроблено інтерактивний дидактичний застосунок, який може використовуватись у викладанні курсу з квантитативної лінгвістики. Результати цієї роботи можуть бути використані у різноманітних сферах: стилістиці, морфології, лексикології, вивченні української мови та власне квантитативної лінгвістики тощо.

Структура роботи: кваліфікаційна робота бакалавра складається із 3 розділів та додатків. Перший розділ – теоретичний, досліджує становлення квантитативної лінгвістики як науки, її основні завдання, сучасний стан досліджень в Україні. У другому розділі описується процес створення програми автоматичного укладання частотних словників слів та частин мови за текстовими матеріалами медійного, художнього, офіційно-ділового і наукового стилів. Також проводиться стилеметричне дослідження на основі цих частотних словників. У третьому розділі наводиться приклад практичного використання матеріалів дослідження: інтерактивний застосунок, який було створено для автоматичного генерування задач із квантитативної лінгвістики. Усі результати роботи наведено у додатках.

Список використаних джерел: 40 позицій, з них 8 англомовних. Також робота містить список із 12 довідкових матеріалів, до якої увійшли документації використаних мов програмування та окремих бібліотек, а також інших електронних інструментів.

Розділ 1. КВАНТИТАТИВНА ЛІНГВІСТИКА ЯК МАТЕМАТИЧНА НАУКА ПРО МОВУ

1.1. Застосування та значущість математичних методів у лінгвістиці

Мова – це надзвичайно складна система дискретних одиниць, які, як і всякі дискретні одиниці будь-якої природи, можуть мати кількісні характеристики. Ці кількісні характеристики властиві одиницям усіх рівнів мовної системи, причому кількісні характеристики одиниць нижчого рівня можуть призводити до якісних відмінностей на вищому рівні. На підтвердження цієї думки можна як приклад навести розбіжність між числом слів (від 10 000 і вище), числом морфем (кілька тисяч), числом складів (від кількох сотень до кількох тисяч) і числом фонем (від 10 до 80). Ці співвідношення зв'язані зі структурою людської пам'яті [Перебийніс 2002, с. 3]. Наведені вище кількісні співвідношення між словами, морфемами, складами і фонемами дозволяють здійснити класифікацію мов, яку можна використати і при вивченні їх історії (наприклад, мови з односкладовими словами-морфемами – це мови з музичним наголосом; дослідження зв’язку між кількістю фонем і середньою довжиною морфеми). Справді, наявність у мові кількісних характеристик та ефективність їх аналізу для різноманітних мовознавчих досліджень повністю визнані.

Маючи статистичні дані для мови, яка є предметом певного дослідження, результати такого дослідження одержать певну авторитетність, адже базуватимуться на конкретних даних, а не умовних припущеннях. Теоретичне значення застосування статистичних методів та їх результатів для мовознавства важко переоцінити. Статистичні методи не тільки додають більшої ваги і авторитетності мовознавчим висновкам, вони здатні розкрити такі закономірності будови мови та мовлення, які без них розкрити неможливо, перевірити і, можливо, навіть відкинути деякі загальноприйняті твердження [Перебийніс 2002, с. 12]. Як зауважив німецький лінгвіст Ґ. Альтманн, «Шлях дисципліни вглиб рано чи пізно наштовхується неминуче на обмеженість якісних методів, на безпорадність неточного способу вираження, на відсутність гіпотез, а також на відсутність теорії» [Альтманн 2005, с. 6].

Отже, кількісні характеристики, властиві одиницям мови і власне мові, можна вважати підставою для підтвердження підпорядкованості мови статистичним законам. Незважаючи на це, мову не можна назвати повністю статистичним явищем, оскільки далеко не всякі кількісні характеристики є характеристиками статистичними.

Захищаючи думку про необхідність статистичних досліджень при вивченні мовних і мовленнєвих явищ, варто сказати, що мова, як суспільне явище, виникає, розвивається та живе у суспільстві, оскільки є засобом спілкування людей, та підлаштовується під потреби того суспільства, у якому вона функціонує [Кочерган 2001, с. 19]. Різні люди мислять по-різному, і у однакових ситуаціях можуть обирати різні мовні одиниці для висловлення спільної думки. При цьому є певне середнє значення, яке називається нормою, а вона встановлюється тільки статистично. Мова розвивається і функціонує у суспільстві, на неї впливає маса індивідів та маса екстралінгвістичних факторів: різні групи і класи суспільства, історичні умови існування народу-носія мови, його зовнішні контакти, ступінь розвитку культури, літератури та ін. Результати впливу цих факторів на розвиток мови не можна завбачити не тільки тому, що їх багато і вони різноманітні, але й тому, що взаємодія їх зі структурними особливостями мови надзвичайно складна і ще не пізнана. Всі спроби впливати на мову за своїм бажанням і розсудом у більшості випадків закінчилися невдачею: мова сприймає невелику кількість нав'язуваних їй правил і одиниць, очевидно, тільки ті, які узгоджуються з її структурою. Тому взаємодія мови і суспільства підкоряється статистичному закону (або законам), і вивчати її треба статистичними методами.

Можна по-різному підходити до питання дослідження мовних та мовленнєвих явищ. Якісний аналіз мови передбачає її категоризацію, тобто виділення в мові певних класів явищ, що об'єднуються за певними якісними ознаками. Однак навіть така категоризація тісно пов'язана з квантифікацією мови, тобто її кількісним аналізом. Таким чином, стає зрозумілим, що мова має

як якісні, так і кількісні характеристики. Більше того, письмове втілення мови – текст – ще більшою мірою демонструє кількісні ознаки.

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. На підтвердження цього сучасна лінгвістика, багата своїми дослідницькими методами та способами аналізу тексту, тяжіє до застосування квантитативних методів у вивченні лінгвальних явищ [Кульчицький 2021, с. 74]. Усі вони об’єднані в одній науці – квантитативній лінгвістиці, яка, у свою чергу, є розділом математичної лінгвістики.

Квантитативна лінгвістика (КЛ) займається дослідженням процесу вивчення мови, її зміни і сфери застосування, а також структури природних мов. Її кінцева мета — сформулювати закони, за якими функціонує мова і, в результаті, побудувати загальну теорію мови у вигляді сукупності взаємопов'язаних законів функціонування мов. Ця галузь емпірично ґрунтується на результатах мовної статистики, яка може інтерпретуватися як статистика мов або статистика лінгвістичного об'єкта.

На статистичних методах може ґрунтуватися і вивчення функційних стилів мов, більше того, можливо проводити відповідні дослідження на основі особливих форм (параметрів), які мовні закони приймають у текстах різних стилів. У таких випадках КЛ проводить дослідження в стилістиці з метою довести настільки об'єктивно, наскільки це можливо, принаймні в одній області дій існування стилістичного феномена, посилаючись на дію мовного закону. Підрозділ квантитативної лінгвістики, який вивчає статистичні параметри різних стилів, називається стилеметрією [Краснобаєва-Чорна 2018, с. 94].

Лінгвостатистика розглядається і як техніка обробки лінгвістичних даних, і як метод дослідження мови та мовлення, і як концепція, система ідей та уявлень про об'єкт лінгвістичної науки. І якщо техніка квантитативної лінгвістики використовується зараз у багатьох дослідженнях, то методика реалізується тільки тоді, коли дослідник відчуває, що лише за допомогою кількісного підходу можна одержати нові дані або перевірити набуті знання про лінгвістичний об'єкт,

коли дослідник переконаний в імовірнісній природі лінгвістичного об'єкта і ставить перед собою завдання – описати його у кількісних показниках [Перебийніс 1985, с. 37].

Кількісні методи в мовознавстві, за Валентиною Перебийніс, не лише уточнюють результати досліджень у мовленні, а й у мові, дозволяючи отримувати науково обґрунтовані висновки, іноді непередбачувані. Вони також забезпечують вірогідність результатів, розкриваючи властивості мовних одиниць та текстової структури, що без них було б неможливо [Перебийніс 2002, с. 7]. Ефективність квантитативних методів зумовлена кількома причинами: точні кількісні дані завжди перевіряються, можна визначити випадковість або значущість відхилень значень показників, встановити необхідний обсяг дослідницького матеріалу, а також мінімізувати суб’єктивність через формалізацію визначення досліджуваних одиниць. Крім того, такі методи допомагають досягти точних висновків.

Для виявлення відмінностей між стилями та жанрами використовуються статистичні параметри, що дозволяють розуміти функціонування мовних одиниць у різних контекстах. Нові дослідження текстів дозволяють підтверджувати або визначати нові статистичні параметри та закономірності будови текстів та структурних особливостей різних мов. Методи аналізу стилю поділяються на експертні (де текст опрацьовують професійні лінгвісти – експерти) і формальні, останні зокрема базуються на машинному навчанні (байєсівський класифікатор, нейронні мережі, дерева рішень, метод опорних векторів, генетичні алгоритми, метод k-найближчих сусідів) або статистичних дослідженнях (одновимірні: критерій Стьюдента, двосторонній критерій Фішера, хі-квадрат Пірсона; багатовимірні: метод головних компонент, критерій Колмогорова–Смірнова, ланцюги Маркова, статистичний кластерний аналіз тощо). Формальні методи дозволяють порівнювати характеристики текстів, представляючи їх у вигляді векторів параметрів, що об’єктивно відображають характеристики тексту.

1.2. Становлення квантитативної лінгвістики

Статистична лінгвістика як наука виокремилася порівняно недавно, однак кількісні та статистичні методи до мови та мовлення застосовували ще з античних часів. Наприклад, у III ст. до н. е. для творчості Гомера александрійські граматисти вручну підраховували, які слова трапляються всього один раз за твір.

В епоху Середньовіччя з метою узгодження різних текстів та перекладів Святого Письма укладалися повні списки його слів з усіма випадками їх використання у текстах. У XVII ст. з'явилася праця, що аналізує розподіл слів у грецькому перекладі Нового Заповіту методом, який майже не відрізняється від сучасного.

У XIX ст. укладають латинські та грецькі словопокажчики, надбаннями статистичної лінгвістики починає користуватися стенографія – швидкий дослівний запис усного мовлення за допомогою системи спеціальних умовних знаків та скорочень найчастотніших буквосполучень, слів, словосполучень, виразів. Для вдосконалення системи стенографії на матеріалі 11 млн. слів був укладений частотний словник німецької мови Кедінґа, виданий у Берліні 1898 р. [Бук 2008, с. 15].

Новим поштовхом до розвитку статистичної лінгвістики стало зростання популярності вивчення іноземних мов у середині XIX ст. Педагоги дійшли висновку, що доцільно було б створити словники слів, які зустрічаються у текстах частіше за решту, а у 1920 р. Кеністон вперше вказав на те, що важливість слова пов'язана не лише із його частотністю, а й з тим, у якому із функційних стилів воно трапляється. У цій сфері надзвичайно скрупульозну роботу здійснив англійський мовознавець і педагог Палмер, відібравши 3000 слів, які дають змогу розуміти 95% тексту [Палмер 1924].

Історія сучасної статистичної стилістики, або стилеметрії, починається у час, коли англійський математик А. де Морган висловив припущення, що різні автори можуть бути визначені за допомогою прихованих статистичних характеристик [де Морґан 1847]. Застосування математичних і статистичних методів у мовознавстві почалося в середині XIX – на початку XX століть завдяки працям українського математика В. Буняковського, німецьких вчених Е.

Ферстемана та Ф. Кедінґа, а також російського математика А. Маркова. Надалі ці методи були розвинуті в роботах Ґ. Альтмана, Р. Кьолера, П. Ґжибека, Г. Віммера, А. Павловскі, Я. Самбор, Ю. Тулдави, Р. Піотровского, А. Шайкевича та інших.

Перші спроби статистичного аналізу текстів, зокрема підрахунок слововживань у творах різних авторів та порівняння отриманих даних були здійснені на початку XX ст. Прикладом таких досліджень є праця А. Маркова "Приклад статистичного дослідження над текстом «Євгенія Онєгіна»" [Марков 1913] та робота М. Морозова "Лінгвістичні спектри: засіб для відрізнення плагіатів від справжніх творів того чи того відомого автора" [Морозов 1915]. Такі дослідження мали на меті встановлення авторства або плагіату.

З появою комп'ютерів для дослідників відкрилися нові можливості та методи дослідження лінгвістичних явищ. Вдалося максимально зменшити використання механічної роботи, що значно скорочувало час для підрахунків, пошуку, лематизації і сортування слів. У другій половині XX ст. активізувалися лінгвостатистичні дослідження, з’явилися фундаментальні праці, що започаткували нові напрями мовознавства, такі як глотохронологія (М. Сводеш "Лексико-статистичне датування доісторичних етнічних контактів" [Сводеш 1952]); лінгвістична типологія (Дж. Грінберг "Квантитативний підхід до морфологічної типології мов" [Грінберг 1960]); структурно-математична та прикладна лінгвістика. Наприкінці XX ст. з’явилися спроби опису основ квантитативної лінгвістики, такі як праці О. Надточого ("Статистична діагностика авторських відмінностей у синтаксисі", 1983), Е. Вудса, П. Флетчера, А. Х'юза ("Статистика в лінгвістичних дослідженнях", 1996).

Початок XXI ст. відзначився теоретико-прикладними та лексикографічними працями з квантитативної лінгвістики, такими як роботи Ш. Еверта ("Кембриджський словник статистики", 2002), В. Перебийніс ("Статистичні методи для лінгвістів" [Перебийніс 2002]), Н. Дарчук ("Комп'ютерна лінгвістика (автоматичне опрацювання тексту)" [Дарчук 2008]) та інші.

1.3. Сучасний стан та завдання лінгвостатистики

За більше ніж 80 років вона розвивається на межі класичної лінгвістики і пережила всі школи, всі парадигми, що виникали на її шляху. Мабуть тому, що вона у пошуках істини вона не відходила від своїх фундаментальних принципів. Дійсним є твердження, що "кожна достатньо розвинена наукова дисципліна рано чи пізно, принаймні, на певному етапі свого розвитку, може опинитися на порозі математизації" [Альтманн 2005, с. 6], і лінгвістика не стала винятком. Саме статистичні методики з комп'ютерною підтримкою відкривають нові шляхи для дослідження літератури та мови, а також мають неоціненний потенціал для вирішення багатьох теоретичних та практичних завдань лінгвістики й NLP (Natural Language Processing).

Результати, виявлені методами статистичної лінгвістики, плідно застосовують у багатьох сферах сучасної науки: судовій та кримінальній лінгвістиці, лінгводидактиці, психолінгвістиці, дешифруванні історичних писемностей, глоттохронології, соціолінгвістиці, стенографії, стилеметрії, комп'ютерних технологіях [Бук 2008, с. 7]. Також це стосується усіх галузей класичної лінгвістики: фонології, граматики, лексикології, семантики, денотативного аналізу, історичної лінгвістики, аналізу тексту, діалектології, тощо, що дало змогу здійснити відкриття, неможливі без використання математичного підходу. Жодна інша лінгвістична дисципліна не мала такого впливу на інші науки як квантитативна лінгвістика. Закон Ціпфа є предметом щонайменше двадцяти інших дисциплін, які його аналізують і розвивають [Ціпф 1949]. Все більше дослідників цілком відмінних наук, таких як фізика, математика та біологія, підключаються до дослідження мови.

За допомогою квантитативних методів вивчають всі мовні підсистеми. Проводяться спроби систематизувати лінгвостатистичні дослідження з метою виділення галузей мовознавства, де ці методи успішно застосовуються. Досліджуються фонетичний лад, лексичний склад, авторські та функційні стилі, властивості текстів з урахуванням різних параметрів. Аналізуються також різні

кількісні характеристики тексту та мовних одиниць. Досліджуються швидкість та закономірності розвитку мови, проводиться порівняльне вивчення мов та їхніх підсистем, діалектологія, психолінгвістичні експерименти та реконструкція прамови. Квантитативні методи також застосовуються у вивченні семантики мови, перекладознавстві та методиці викладання іноземних мов. У майбутньому разом із розвитком світової лінгвістичної науки можна говорити лише про розширення завдань цієї галузі.

Зараз у лінгвістиці дуже багато дисциплін, залежних від комп'ютера. Їх прогрес, як і прогрес будь-якої науки, гарантований лише тоді, коли всі напрями відкриті для дослідження, а не розглядається лише вузьке коло проблем або береться до уваги лише один аспект. Це далеко не завжди залежить від науковців, адже для розвитку певної діяльності конче необхідні ресурси, як фінансові, так і людські, так і інтелектуальні. На сьогодні українські інститути та лабораторії, які займаються статистичними дослідженнями української мови, використовують потенціал лінгвостатистики не повною мірою. В Україні бракує досліджень рідною мовою на цю тематику, а створені частотні словники, хоч і є неймовірно цінними для досліджень – застарілі й не демонструють стан сучасної мови з тією точністю, що можна досягти на сучасних текстах, зважаючи на введення нового правопису, враховуючи, що «перехідний період» завершується у 2024 році. Систематизація даних за допомогою частотних словників має неабияку практичну цінність, такі словники широко використовуються науковцями для статистичних досліджень мови, а отже, їхня база має постійно поповнюватися для формування точних висновків і про сучасний стан мови, і про її використання у різних сферах нашого життя.

1.4. Статистичні дослідження в українському мовознавстві

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. Спочатку вони стосувалися відбору лексичного мінімуму іноземних мов, пізніше було започатковано планомірне статистичне дослідження українських текстів художнього, науково-технічного та соціально-політичного функційних стилів, виявлено їхні статистичні параметри. Після цього дослідження світ побачили численні монографії та збірники [Бук 2008, с. 16]. Наступний проєкт концентрувався на описі сполучуваності англійських іменників, прикметників та дієслів, його результати були опубліковані в “Довіднику найбільш уживаних англійських словосполучень” за редакцією В. Перебийніс (1986).

Пізніше з метою виявлення лексичних особливостей різних стилів було укладено частотні словники для п‘яти функційних стилів: ЧС сучасної української художньої прози на вибірці текстів 25 письменників загальним обсягом 500 000 слововживань; ЧС публіцистики за текстами всеукраїнських газет за 1994 р.; ЧС розмовно-побутового стилю склали 45 текстових уривків з прозових драм 1982–2002 рр.; ЧС наукового стилю за такими галузями наук: біологія та хімія, психологія і педагогіка, фізика та математика, техніка, географія та геологія, історія та мовознавство; ЧС офіційно-ділового стилю за текстами підручників ділової мови та офіційних документів (Конституція України, кодекси, українські та міжнародні закони та ін.; ЧС поетичної мови за текстами п‘ятнадцяти поетів 1975–1995 рр.. Обсяг останніх ЧС однаковий – 300 000 слововживань. Варто зазначити, що всі вони були складені автоматично, окрім ЧС художньої прози, який збирали вручну [Дарчук 2005, с. 100].

Нині статистичні дослідження проводять в Українському мовно-інформаційному фонді, Лабораторії комп’ютерної лінгвістики Київського національного університету ім. Тараса Шевченка, Львівському національному університеті ім. Івана Франка, Національному університеті «Львівська політехніка» та ін.

Головним центром досліджень у сфері квантитативної лінгвістики в Україні є відділ структурно-математичної лінгвістики Інституту української мови НАН України [6], який очолює професор Є. Карпіловська. Історія цього відділу складається з трьох етапів, кожен з яких відзначається багатовекторними дослідженнями мови: впровадження структурних, статистичних методів і моделювання у вивченні мов; підготовка узагальнювальних праць зі структурної граматики української мови, розробка принципів статистичного аналізу текстів різних стилів, створення частотних словників; розробка систем автоматизованого аналізу текстів, створення комп'ютерного морфемно-словотвірного фонду української мови та нових словників. Сьогодні відділ досліджує інноваційні процеси в сучасній українській мові, їх комп'ютерне моделювання, вплив соціодинаміки на мову та тенденції мовних змін.

Висновки до розділу 1.

Отже, мова, будучи складною системою дискретних одиниць, може бути аналізована за допомогою якісних та кількісних методів. Залежно від поставлених цілей і завдань дослідження мовних явищ лінгвіст може використовувати обидва ці підходи. Однак іноді виникають завдання, які можна вирішити лише за допомогою кількісних методів. Саме тому виникла і бурхливо розвивається квантитативна лінгвістика. Це призвело до впровадження в лінгвістику різноманітних кількісних методів і моделей, що використовуються в природничих та суспільних науках. Цей напрямок сприяє появі нових теоретичних концепцій, розв'язанню практичних проблем у різних сферах лінгвістики та допомагає обґрунтовувати одержані результати реальними даними. При цьому є потреба у ще більшій кількості якісних та детальних досліджень, особливо українцями для української мови, задля глибшого розуміння сучасних та історичних мовних процесів.

РОЗДІЛ 2. ОБЧИСЛЕННЯ СТАТИСТИЧНИХ ПАРАМЕТРІВ ФУНКЦІЙНИХ СТИЛІВ УКРАЇНСЬКОЇ МОВИ

2.1. Організація текстового матеріалу статистичного дослідження

Завдання, які ми ставимо перед собою для цього дослідження – створити програму автоматичного укладання частотних словників слів та частин мови, щоб на базі цих ЧС здійснити стилеметричне дослідження. Для цього нам необхідно визначити текстові бази та розмір вибірок стилів української мови, за якими ми побудуємо частотні словники. Після цього ми здійснимо попередню автоматичну обробку текстових файлів – "почистимо" їх від зайвих символів та токенізуємо, проведемо лематизацію всіх словоформ та припишемо їм відповідний тег частини мови. Обов'язковий етап роботи – поділ кожної вибірки на підвибірки обраного розміру. Тільки тоді ми зможемо укласти необхідні частотні словники та обчислити статистичні характеристики частин мови у різних функційних стилях. Опишемо детальніше кожен етап роботи.

Робити висновки про обрані стилі: медійний, художній, офіційно-діловий та науковий ми будемо робити на основі відповідних текстових вибірок, оскільки генеральна сукупність для кожного стилю необмежена і постійно поповнюється новим матеріалом. Для якомога точнішого та ефективнішого аналізу ознак функційних стилів було складено чотири вибірки текстів українською мовою. Детально опишемо кожну із них.

Законодавчі тексти. До цієї вибірки увійшли закони, заяви, рішення тощо, видані у грудні 2023 року державними інституціями, такими як Офіс Президента, Національний Банк, Кабінет Міністрів та ін. Ці тексти повною мірою відображають офіційно-діловий стиль сучасної української мови. Усіх правил однорідності дотримано. Обсяг вибірки – 60 490 токенів.

Художні тексти. До цієї вибірки ми внесли два романи сучасної української письменниці Євгенії Кузнєцової – «Драбина» [Кузнєцова 2023] та «Спитайте Мієчку» [Кузнєцова 2021]. Правил однорідності було дотримано. Загальний обсяг – 52 505 слововживань.

Наукові тексти. Ця вибірка вміщує наукові статті та праці на теми дерматології та суміжних галузей. Хронологічно тексти обмежені 2023 роком. Обсяг – 57 523 слововживань, адже деякі з праць містять численні розрахунки та оригінальні терміни англійською або латинською мовами.

Медійні тексти. До цієї вибірки увійшли новинні статті авторів електронного видання “Українська правда” [6] на військово-політичні теми. Варто зауважити, що видання має свій особливий стиль публікацій, які, проте, сміливо можна вважати наочним прикладом якісної роботи журналістів. Хронологічні межі текстів – жовтень 2023 – травень 2024 року. Розмір вибірки – 53 004 слововживань.

Текстовий матеріал дослідження організовувався з дотриманням вимог лінгвістичної та статистичної репрезентативності текстових вибірок.

Задля лінгвістичної однорідності усі тексти вибірки обмежені хронологічно, жанрово та тематично, а також належать авторам зі схожою манерою письма. Але обов’язковим принципом нашого дослідження є актуальність, тож хронологічно тексти обмежені 2021-2024 роками створення та публікації, що дозволяє одержати дані про сучасну українську мову для подальшої роботи. Також задля того, щоб тексти репрезентували саме стан сучасної української мови, ми слідкували за тим, щоб до файлів не потрапили перекладені тексти, хоч ми не заперечуємо того, що переклади теж відображають вигляд мови на певному етапі її розвитку.

Вибірка репрезентативна тоді, коли вона рівномірно розподіляється по генеральній сукупності, а не відображає лише невелику її частину, та має досить великий обсяг для вірогідних висновків про властивості цілої генеральної сукупності.

Залежно від методів організації, розрізняють механічну, випадкову та типову, або зональну, вибірки [Перебийніс 2002, с. 17]. Оскільки типова вибірка є найрелевантнішою в лінгвістичних дослідженнях, особливо при дослідженні мовленнєвого матеріалу, при якому не можна нехтувати існуванням різних типів і видів функціонування мови, то під час дослідження ми складали вибірку саме

за цим принципом. Генеральна сукупність за певними лінгвістичними або лінгвостилістичними ознаками ділиться на зони. Зоною може слугувати множина текстів, слів або інших одиниць мови. Найважливішим є виділення критеріїв, за якими обрані зони встановлюються, після чого відбір одиниць, як правило, здійснюється випадковим чином. Ці критерії ми описали раніше.

Для того, щоб усі частотні словники були однаковими за об’ємом, розмір кожної вибірки складатиме 50 000 слововживань. Такий обсяг достатній для вірогідних висновків про властивості цілої генеральної сукупності. Проте довжина зібраних файлів більша. Це пов’язано з тим, що для створення ЧС нам потрібні лише слова українською мовою, а обрані тексти часто містять числа та слова, написані латинським алфавітом (наприклад, для назв речовин, організацій, видів зброї тощо), і завжди містять розділові знаки, тож файл проходитиме етап “чистки” від непотрібних символів. З метою дотримання запланованої величини вибірки незалежно від кількості видалених одиниць ми зібрали більше текстів, аби пізніше автоматично уніфікувати розміри частотних словників.

2.2. Автоматичне укладання частотних словників

Частини мови – великі класи слів, об'єднаних спільним загальним граматичним значенням і його формальними показниками. І. Вихованець надає таку дефініцію цього поняття: «найзагальніші семантико-граматичні класи слів, що характеризуються такими чотирма ознаками: 1) узагальненим (категорійним) граматичним значенням, абстрагованим від конкретних лексичних значень слів; 2) структурою граматичних категорій; 3) системою форм словозміни або її відсутністю; 4) спільністю синтаксичних функцій» [Вихованець 2004, с. 11].

Розподіл слів за частинами мови – досить складне завдання, й існуючі класифікації не завжди задовольняють дослідників. Справа в тому, що принципи віднесення слів до різних частин мови досі не усталені, і в різних граматичних студіях визначається неоднакова кількість частин мови. Відрізняються також погляди на критерії класифікації, їх кількість та якість. Одні віддають перевагу

лексичному значенню слова, інші - його формальній структурі. Наприклад, О. Безпояско, К. Городенська та В. Русанівський використовували класифікацію, в основу якої покладено морфологічний принцип, що доповнюється синтаксичним і лексико-семантичним. Таким чином, вони виділили дев'ять частин мови – іменник, дієслово, прикметник, числівник, прислівник, прийменник, сполучник, частку та вигук, вважаючи, що займенник "не виражає власного лексичного значення, а лише опосередковано представляє щось, узагальнено вказуючи на предмет, ознаки, ознаки ознак", а отже, не є окремою частиною мови [Безпояско 1993]. І. Кучеренко виділив сім частин мови – іменник, прикметник, числівник, дієслово, прислівник, частка, сполучник – та обґрунтував безпідставність виділення прийменників як окремої частини мови; щодо них він уживав термін прийменникові слова та розглядав у складі прислівників узагальненого значення [Кучеренко 2003]. І. Вихованець запропонував іншу оригінальну систему поділу слів за частинами мови, згідно з якою центральними частинами мови вважаються іменники і дієслова та периферійними – прикметники і прислівники, а усі інші традиційно відокремлювані класи слів можна пояснити і замінити іншими термінами, які відображають їхню суть [Вихованець 2004]. Тих самих поглядів дотримується й А. Загнітко. І це далеко не повний список можливих класифікацій. За різними критеріями в українській мові виділяють від чотирьох до чотирнадцяти частин мови.

Під час виконання цієї роботи ми використовуватимемо традиційну систему частин мови. І оскільки обраний морфологізатор pymorphy3 [35] для української мови виділяє 18 класів, виникла потреба звести деякі з них під спільний тег за такою схемою:

- NOUN – іменник
- VERB (verb, infn, prtf, prts, grnd) – дієслово
- ADJ (adjf, adjs) – прикметник
- ADV (advb, comp, pred) – прислівник
- NPRO – займенник
- NUMR – числівник

- PREP – прийменник
- CONJ – сполучник
- PRCL – частка
- INTJ – вигук

Зібравши необхідний для дослідження матеріал та обґрунтувавши використану систему частин мови, можемо переходити до написання програмного коду для створення частотних словників за чотирма вибірками. Першим і обов'язковим кроком для подальшого опрацювання інформації є конвертування текстових файлів з формату docx у формат txt. Готовий код та конвертовані файли знаходяться на прикріпленому до курсової роботи дискові (див. Додаток 1).

Тепер можемо створити концепцію частотних словників. Принципи укладання абсолютно однакові для усіх чотирьох. Одиницею підрахунку виступатиме слово, джерела – описані вище вибірки текстів визначених функційних стилів обсягом 50 000 слововживань кожна. Для стилеметричного дослідження, яке ми проведемо пізніше, нам необхідно кожну вибірку поділити на підвибірки оптимальним обсягом 1000 слововживань, усього для кожного стилю буде 50 підвибірок. Граматична інформація, яку представлятимуть ЧС, складається зі словоформи, витягнутої з тексту, відповідної леми та тегу частини мови. Статистичні дані включають абсолютну частоту одиниці у вибірці та розподіл частоти одиниці у підвибірках. Усі ЧС укладатимуться автоматично за допомогою програмного коду. Як інструмент візуалізації ми будемо використовувати DB Browser for SQLite [DB Browser 2025], у якому можна автоматично фільтрувати бази даних, тож немає потреби створювати окремі словники за спаданням/зростанням частоти, алфавітним порядком чи частиномовною належністю.

За допомогою бібліотеки для роботи з регулярними виразами re [re 2025] кожен конвертований файл очищуємо від усіх розділових знаків, цифр та латинських літер, які можуть зустрічатися у тексті.

Оскільки одиницею підрахунку наших ЧС є слово, потрібно ввести поняття токенізації – процесу розбиття тексту на менші одиниці (у нашому випадку слово) та лематизації – зведення словоформи, яка функціонує в тексті, до її початкової, словникової форми. Токенізацію ми виконуватимемо за допомогою бібліотеки tokenize_uk [tokenize_uk 2025], а лематизацію – pymorphy3 [pymorphy3 2025]. Останню бібліотеку використовуємо також для частиномовної розмітки тексту.

Основною проблемою правильного виконання цієї операції є лексична та граматична омонімія. Інколи, але досить рідко, розрізняють окремі значення багатозначних слів, проте повністю автоматично цього сьогодні зробити неможливо. Деякі поширені помилки ми прокоментуємо дещо пізніше. Ще однією перешкодою є відсоток слів, яких морфологізатор не розпізнав, а отже, не приписав жодного тегу частини мови. З одного боку, таке явище створює небажану похибку в дослідженні, бо замість “невизначених” слів ми могли б аналізувати чисті та точні результати. З іншого боку, це можна розглядати як ще одну статистичну характеристику стилю, а також додатково проаналізувати список цих токенів.

Отже, було створено вісім частотних словників:

- ЧС словоформ медійного стилю
- ЧС частин мови медійного стилю
- ЧС словоформ художнього стилю
- ЧС частин мови художнього стилю
- ЧС словоформ наукового стилю
- ЧС частин мови наукового стилю
- ЧС словоформ офіційно-ділового стилю
- ЧС частин мови офіційно-ділового стилю

Словоформи та частини мови розташовуємо за спадом частот. Перевіримо, чи справді заданого раніше запасу слів вистачило для одержання бажаного обсягу словоформ у вибірках. Тепер, коли частотні словники готові до перегляду, переходимо до їх детального аналізу.

Як видно із ЧС (див. Додаток 2-3), мовлення надає перевагу невеликій кількості одиниць, які використовуються помітно частіше за решту. Вони становлять ядро мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними. Цю закономірність зауважив ще Дж. Дьюї на початку XX ст. та назвав її законом переваги. Детальніше цей феномен дослідив німецький мовознавець Дж. Ціпф, сформулювавши закон, який встановлює залежності частоти слова у мовленні та словнику: чим частотніше слово, тим вище воно стоятиме у ЧС [Ціпф 1949].

Це помітно і в наших ЧС, адже лише кілька десятків слів мають абсолютну частоту вище 100 (зазвичай це приблизно 0,5% слів списку), тоді як слова з частотами 1 та 2 займають в середньому 60-80% списку залежно від стилю. Звернімо увагу, що найнижчу частку низькочастотних слів має офіційно-діловий стиль (58%), а найвищу – художній (81%). Верхні позиції усіх списків очікувано займають сполучники, частки, прийменники та займенники. Повертаючися до закону Ціпфа, варто нагадати, що у кожному ЧС його варто застосовувати до відповідного функційного стилю, а не української мови загалом.

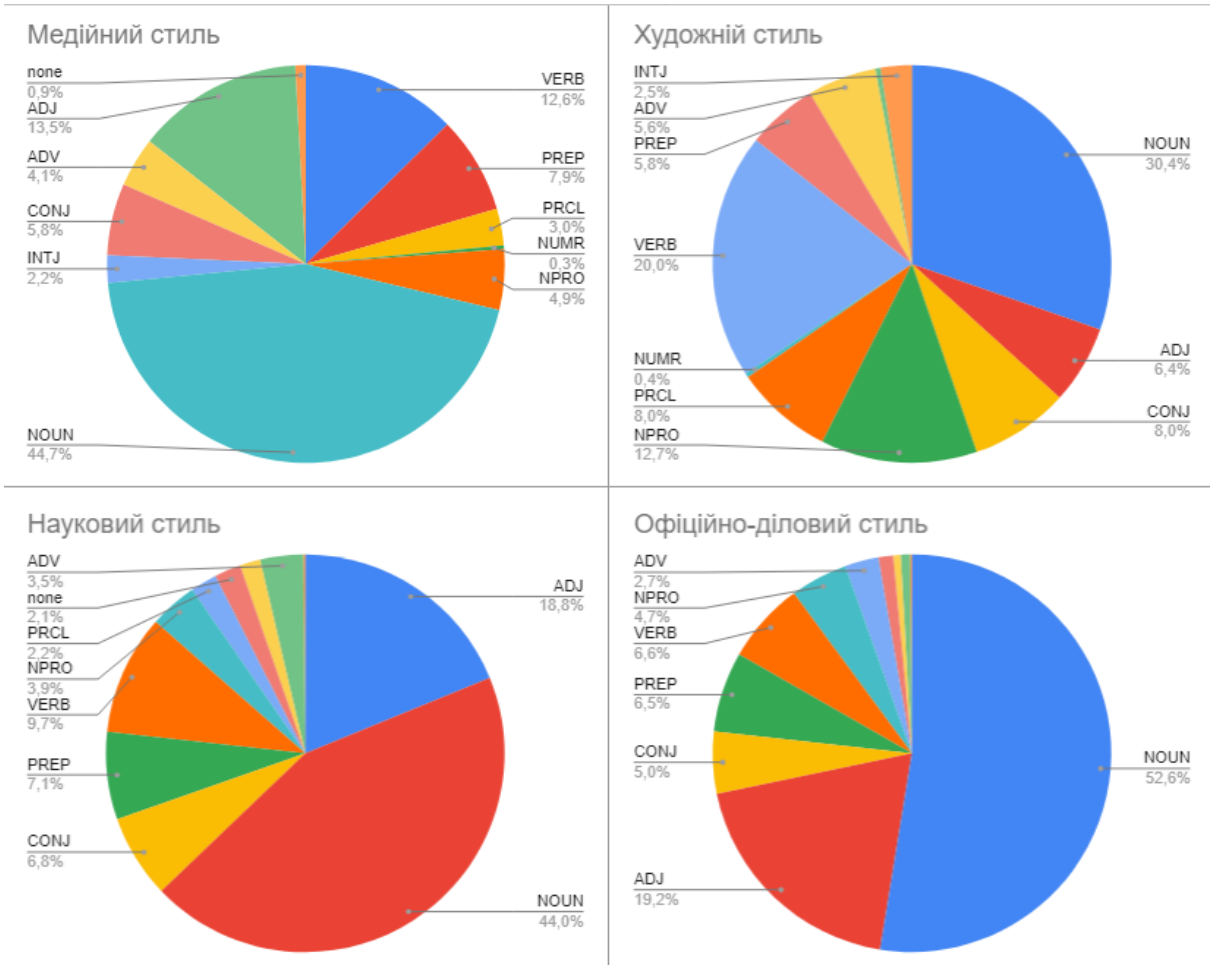
На жаль, виконана автоматично робота, хоч і значно швидше дозволяє отримати результати, має свою похибку. Вище ми вже описали дві основні проблеми проміжних процесів лематування та частиномовної розмітки, і зараз дослідимо їх вплив на точність створених частотних словників.

Ми перевірили 100 найчастотніших слів у кожному ЧС, і знайшли спільні для усіх списків помилки. Наприклад, прийменник *на* класифікується як вигук (у значенні *тримай*), слово *що,* яке залежно від місця у реченні може належати до різних частин мови, зводиться до сполучника, прийменники *до* та *про* позначені як іменники (перший, очевидно, від ноти *до*), до часток віднесли слова *як, ще, коли* та *це*. Було знайдено багато прикладів помилкового тегу саме через неправильно підібрану лему: його – йога – NOUN, іруся – іертися – VERB, була – булий – ADJ, протягом – протяг – NOUN, при – перти – VERB, бути – бута – NOUN, тому – том – NOUN та безліч інших. Важливо те, що усі ці слова мають високу частоту вживання у текстах, і помилка складає не кілька одиниць, а

десятки і подекуди навіть сотні одиниць, що негативно впливає на точність одержаних даних. Вручну цієї проблеми виправлено не було, оскільки з кожним новим запуском програмного коду рядки з помилками знову з'являються.

Щодо списку слів без тега частини мови, до них увійшли нерозпізнані скорочення та абревіатури, деякі вставні слова (це може бути пов'язано з тим, що багато дослідників дійсно не відносять їх до жодної частини мови, а вважають окремим розрядом слів), деякі пестливі форми слів. Цікавий цей список у частотному словнику наукового стилю – у працях подекуди застосовувалися формули з використанням літер грецького алфавіту, а оскільки їх кодифікація відмінна від латинських символів, файли від них не чистилися, і математичні змінні потрапили до основного тексту. Крім цього, часто траплялися слова з одруківками або набір літер, який не має жодного відомого значення. Отже, у текстах медійного стилю було 93 слова без тегу частини мови, у текстах художнього – 49, наукового – 238, офіційно-ділового – 18.

Важливий показник, на який варто звернути увагу при перегляді частотних словників частин мов (див. додаток 2,3) – частка кожної категорії у вибірці. Порівнявши їх для різних вибірок, можна отримати дуже цікаві відомості про функційні стилі української мови, які візуалізовано графічно на діаграмах мал. 2.1. Наприклад, іменник має найвищу відносну частоту в офіційно-діловому стилі (52,6%), а найнижчу – художній (30,4%), тоді як для медійного та наукового частка складає 44,7% та 44% відповідно. Для офіційно-ділового стилю особливістю також є досить мала частка дієслів (усього 6,6%) та порівняно значна частка прикметників (19,2%), останнє також притаманне науковим текстам (18,8%). Вдвічі і навіть втричі більше займенників використовують у художній прозі (12,7%), п'яту частину вибірки становлять дієслова, порівняно мала частка прикметників (6,4%). Подібний розподіл відносної частоти мають такі службові частини мови: сполучник, прийменник (5-8%, з мінімальним та максимальним значеннями у законодавчих та художніх текстах відповідно), а ось відсоток частки доволі варіюється – від 8% для художнього стилю та менше 1% для офіційно ділового.



мал. 2.1. Розподіл частин мови у функційних стилях

2.3. Обчислення основних статистичних характеристик вибірок

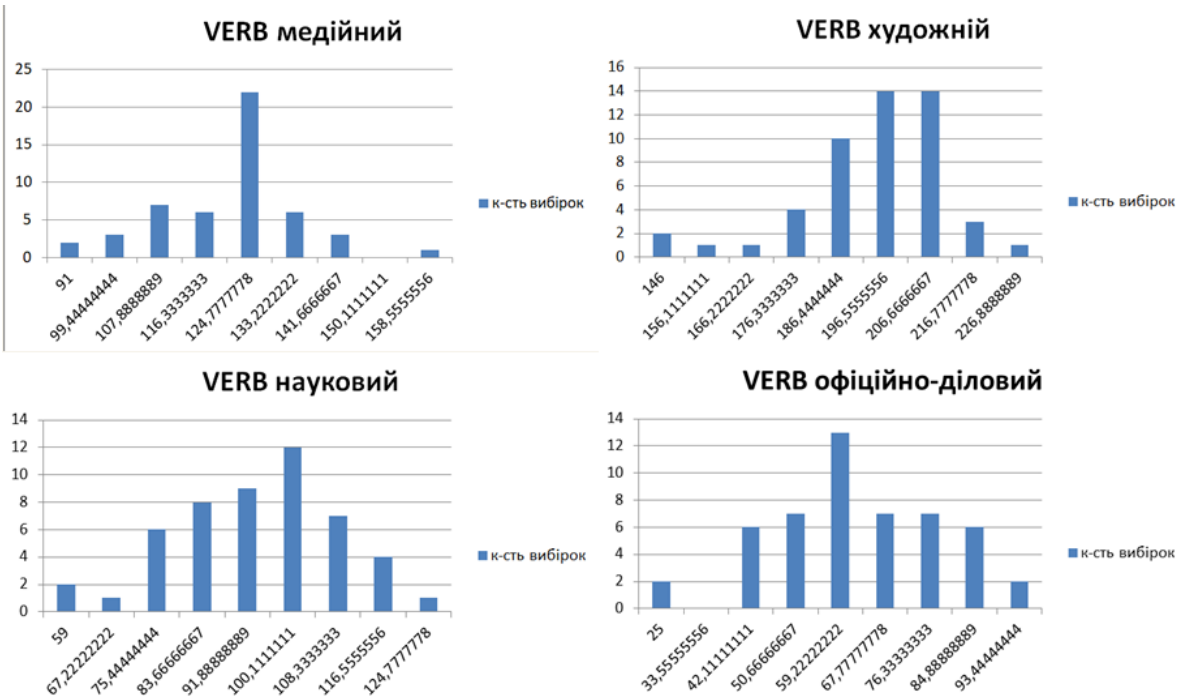
Після того як ми організували наші вибірки та створили частотні словники на їхньому матеріалі, варто визначити величини підвибірок. Оскільки розмір вибірки становить 50 000 токенів, для отримання наочних результатів та цінних висновків дослідження, ми поділимо текст на 50 підвибірок по 1000 (див. Додаток 2, 3) токенів кожна. Це також робимо автоматично (див. Додаток 4).

Тепер починається наступний етап статистичного дослідження — визначення основних статистичних характеристик частин мови, тобто тих кількісних показників, які характеризують поведінку досліджуваного явища в конкретній вибірці. Це обчислення також включає в себе підрахунок абсолютних частот одиниці в кожній підвибірці, обчислення середньої частоти досліджуваної одиниці, знаходження середньої частоти, середнього квадратичного відхилення, міри коливання середньої частоти, коефіцієнтів варіації та стабільності.

Зупинимося детальніше на кожному значенні та спробуємо порівняти деякі з них у межах однієї та кількох вибірок.

Варіанти абсолютної частоти. Для усіх вибірок спостерігається логічна тенденція: чим вища абсолютна частота частини мови, тим вищі варіанти частот по вибірках. Якщо побудувати варіаційний ряд (див. Додаток 2), можна помітити, що різниця між максимальним та мінімальним значеннями теж прямо пропорційна варіантам частоти.

Середня частота. Цей показник також є вищим для значень із вищими абсолютними частотами. Якщо порівнювати його з попереднім значенням, то варто звернути увагу, що одна варіанта зустрічається частіше у кількох вибірках, якщо вона ближча за значенням до середньої частоти. Це стосується усіх частин мови і чудово представляється інтервальними рядами. Їх ми побудували за допомогою Excel [Excel 2025] (мал. 2.2.). Інтервальний ряд розподілу – це ряд, де значення варіант подані у вигляді інтервалів, а частоти відносяться не до окремої варіанти абсолютної частоти, як у варіаційних рядах, а до всього інтервалу.



мал. 2.2. Полігони частот для інтервальних рядів дієслів.

Середнє квадратичне відхилення. Це значення показує, на яку величину можуть відхилятися від середньої частоти абсолютні частоти. Для більшості

частин мови середнє квадратичне відхилення варіювалося лише на кілька одиниць. Цікавими є деякі результати для іменника та прикметника. Значення середнього квадратичного відхилення для законодавчих текстів вдвічі перевищило таке ж для художніх текстів у іменнику (для порівняння: 19,58 для художнього стилю, 30,69 для медійного, 38,17 для офіційно-ділового, 36,47 для наукового), а для прикметника аж утричі (для порівняння: 15,89 для художнього стилю, 43,96 для офіційно-ділового, 23,52 для медійного, 22,98 для наукового). Це свідчить про те, що у художній прозі частини мови ведуть себе стабільніше і не відхиляються від норми на великі значення. Таке саме явище можемо спостерігати і на побудованих інтервальних рядах. Порівняти значення середнього квадратичного відхилення для усіх частин мови можна за допомогою створених таблиць (див. Додаток 2, 5).

Міра коливання середньої частоти показує, наскільки може відхилятися значення середньої частоти при зміні вибірки в генеральній сукупності. Ці можливі зміни, що допускаються законами статистики, вважаються статистично неістотними, а тексти, що характеризуються середніми частотами, відмінність між якими статистично неістотна, можна об'єднати в одну вибірку і вважати статистично однорідними. Та в нашому дослідженні довірчі інтервали не перетинаються, за винятком поодиноких випадків (наприклад, середні частоти іменника для медійного та наукового стилю потрапляють у проміжок 443,14-445,56 з ймовірністю 68,3%, і це єдиний випадок для інтервалу з цією ймовірністю), тож вибірки дійсно належать до різних функційних стилів і не можуть бути об’єднані між собою.

Коефіцієнт варіації показує, яку долю середньої частоти складає середнє квадратичне відхилення та вимірює стабільність розподілу однієї одиниці у різних вибірках. Загальна тенденція така: чим більша абсолютна частота одиниці у вибірці, тим нижчий коефіцієнт варіації, тим стабільніша, у висновку, одиниця. Найстабільнішою для всіх вибірок частиною мови виявився іменник, менш стабільно зустрічаються займенник та прикметник, в усіх вибірках, окрім законодавчих текстів, доволі стабільне дієслово. За усіма даними

найстабільнішою вибіркою є художня проза, адже ми узяли твори однієї авторки. Найменш стабільні законодавчі тексти, хоч офіційно-діловий стиль підчиняється певним регламентам, тож можна припустити, що до вибірки просто увійшли різні типи документів. **Коефіцієнт стабільності** тісно пов'язаний з коефіцієнтом варіації, але подає стабільність розподілу одиниці у відсотковому співвідношенні, зручному для сприйняття.

Приклади результатів роботи програми подані у додатку 5.

2.4. Зіставлення статистичних характеристик різних вибірок

Визначення статистичних характеристик певної вибірки — лише перший етап статистичного дослідження. Наша мета полягає не в тому, щоб одержати ряд чисел, а в тому, щоб на їх підставі здійснити змістовні висновки, які дозволяють вивчити глибинну природу кожного функційного стилю. Зробити певні лінгвістичні висновки про характер функціонування мовних одиниць і категорій у мовленні неможливо без якісного зіставлення обчислених даних за допомогою математичних методів. Тобто аналіз лише однієї вибірки ізольовано від інших не дає підстав сказати, що вона належить до іншої генеральної сукупності.

Аналіз результатів попереднього розділу повинен показати, що є спільним і різним у розподілі частин мови у текстах різних функційних стилів. Та це дає лише наочну інформацію про окрему вибірку, і, порівнюючи між собою полігони частот для інтервальних рядів або коефіцієнти стабільності, ми не можемо точно сказати, наскільки між собою відрізняються ці вибірки і чи взагалі така різниця статистично значима.

Для відповіді на це питання необхідно перевірити так звану нульову гіпотезу, яка формулюється щоразу, коли зіставляються статистичні характеристики двох або більше вибірок. Формулюється вона так: розходження між частотами у порівнюваних вибірках не істотне (статистично не значиме), воно обумовлено статистичними законами, а не характером досліджуваних текстів. Якщо за підрахунками розходження не значиме, то нульова гіпотеза

приймається, що свідчить про те, що зіставлювані вибірки належать до однієї генеральної сукупності, і їх можна об’єднати. Інакше слід шукати причину розходження частот у вибірках [Перебийніс 2002, с. 11].

Для зіставлення обчислених значень ми користуватимемося критеріями однорідності хі-квадрат та критерієм Стьюдента.

Ми обчислили значення хі-квадрат для зіставлення абсолютних частот для кожної частини мови. Результати наведені в таблиці 2.1. Обчисливши значення за формулою, ми обчислили кількість ступенів свободи (147 для усіх) та визначили критичне значення для довірчих ймовірностей 95% (67,5) та 99% (76,2). Оскільки обчислений показник вищий за критичний для кожної частини мови, нульова гіпотеза не приймається в жодному випадку. Тим не менш, ми можемо порівняти, за якими частинами мови вибірки функційних стилів відрізняються більшою чи меншою мірою. Найвище значення має прикметник (823,12) та частка (584,25), найнижче – числівник (286,67) та прийменник (293,37).

таблиця 2.1. Значення Хі-квадрат

Частина мови	Значення Хі-квадрат
Іменник	307,79
Дієслово	341,16
Прикметник	823,12
Прислівник	376,45
Прийменник	293,37
Сполучник	309,68
Займенник	486,14
Частка	584,25
Вигук	293,07
Числівник	286,67

Критерій Стьюдента ми обчислили лише для тих значень, які показували між собою більшу різницю. Усі обчислені значення можна переглянути у таблиці 2.2., де другий та п’ятий стовпчик відповідають назвам стилів двох вибірок, які ми зіставляємо, перший стовпчик відповідає частині мови, для якої ми з цих двох вибірок беремо необхідну кількісну інформацію для обчислення критерію Стьюдента. Маючи значення критерію Стьюдента, можна обчислити відносне розходження, або ступінь розходження, відстань між зіставлюваними масивами. Критерій Стьюдента ми позначаємо як *t*, а ступінь розходження – *l*. За таблицею критичних значень для 98 ступенів свободи довірчі ймовірності складають 1,99 для 95%, 2, 64 для 99% та 3,42 для 99,9%. Усі обчислені значення вийшли значно вищими, тож нульова гіпотеза не приймається, і ми можемо обчислити ступінь розходження для різних частин мови. Ступінь розходження у більшості випадків був високим, що свідчить про велику різницю між зіставлюваними текстами та їх належність до різних функційних стилів.

таблиця. 2.2. Значення критеріїв Стьюдента (t) та ступенів розходження (l)

Частина мови	Вибірка №1	*t*	*l*	Вибірка №2
Іменник	худ	36,25	0,95	оф-діл
Дієслово	худ	39,87	0,95	оф-діл
Дієслово	худ	31,79	0,93	наук
Прикметник	худ	31,73	0,94	наук
Прислівник	оф-діл	8,23	0,76	мед
Прислівник	оф-діл	15,94	0,88	худ
Прийменник	худ	10,57	0,81	мед
Займенник	худ	33,07	0,94	наук
Частка	худ	27,23	0,93	наук
Частка	худ	49,8	0,96	оф-діл

Висновки до розділу 2.

Отже, ми зібрали вибірки для чотирьох стилів української мови: художнього, медійного, офіційно-ділового та наукового, створили частотні словники слів і частин мови для кожної з них. На основі отриманих даних ми змогли обчислити статистичні характеристики для усіх частин мови, візуалізувати деякі з них та за допомогою математичних методів порівняти їхнє розходження. В результаті було визначено, що найбільше відрізняються один від одного художній та офіційно-діловий стиль, тоді як науковий та медійний стиль мають більшу подібність, хоч їх і не можна об’єднувати в одну генеральну сукупність. Одержані стилеметричні дані можна використовувати для характеристики медійного, художнього, офіційно-ділового і наукового стилів української мови, а також для розробки завдання автоматичного визначення стилю тексту.

РОЗДІЛ 3. СТВОРЕННЯ ІНТЕРАКТИВНОГО ЗАСТОСУНКУ ДЛЯ ПРАКТИЧНИХ ЗАНЯТЬ ІЗ КУРСУ “КВАНТИТАТИВНА ЛІНГВІСТИКА”

3.1. Концепція проєкту та формування інфологічної моделі дидактичної системи

Провівши власний дослід, ми зіткнулися з прикладом індивідуальної обробки теоретичних знань на практиці, а отже, неодмінно вивчили щось нове та поглибили ці знання на індивідуальному рівні. Саме тому якісне навчання завжди включає в себе практичний компонент, де студент матиме особистий досвід взаємодії з тим чи іншим явищем. Глибинне вивчення квантитативної лінгвістики, та й подальший розвиток цієї науки напряму залежить від практичної роботи з мовою, яку ми досліджуємо.

З цією метою ми створили застосунок, з допомогою якого студенти матимуть можливість розв’язувати задачі з квантитативної лінгвістики для якіснішого та швидшого опанування мовних та мовленнєвих явищ. Уже давно були видані підручники, які мають цей практичний блок, та створений нами продукт, по-перше, електронний та доступніший порівняно з паперовими посібниками, по-друге, має низку переваг у своєму функціоналі. Варто зауважити, що цей застосунок виник із конкретного запиту на схожий продукт, а отже, ми можемо бути переконані, що студенти та викладачі будуть використовувати надбання цієї кваліфікаційної роботи.

Створення нового продукту – це тривалий процес, який має чітку структуру. Було розроблено декілька підходів до цього завдання, але усі вони мають схожий принцип. Для нашої програми ми візьмемо за основу сучасні тенденції розробки. New Product Development – це концепція, що передбачає виведення нового продукту на ринок за допомогою основних ринкових принципів. Іншими словами, це процес перетворення потреб ринку на ідеї, які реалізуються за допомогою готового продукту [40].

Її можна поділити на 7 основних етапів: генерація ідеї, перевірка ідеї, розробка концепції та тестування, бізнес-аналіз, розробка продукту, тестування

на вихід на ринок. Оскільки наш продукт некомерційний і ми не плануємо для нього жодних фінансових витрат чи прибутку, ми пропустимо етап бізнес-аналізу, або ринкової стратегії.

Створення ідеї – ключовий і дуже динамічний фактор. Проблема у тому, що більшість згенерованих ідей не проходять наступний етап – перевірку, часто відкидаються, часто змінюються та додаються в ході роботи за рахунок нових ідей, які сприяють покращенню фінального результату. Буває, що реалізована ідея ніколи не побачить світ або виявиться неактуальною для користувачів. Для роботи над хорошою довершеною ідеєю важливо враховувати кілька точок зору, відгуки та зворотний зв'язок від інших учасників проєкту та зацікавлених осіб.

Ми вирішили зосередитися на проблемах та завданнях потенційних користувачів. Їхня залученість – це можливість швидко згенерувати ідею та створити продукт із високим рівнем затребуваності на ринку. Тобто ми робимо одразу те, що необхідно користувачам.

Ідея про створення електронного збірника задач виникла задовго до початку написання кваліфікаційної роботи, ще у червні 2024 року, коли було отримано запит від гаранта освітньої програми "Прикладна (комп'ютерна) лінгвістика та англійська мова" Зубань Оксани Миколаївни. Студенти мають можливість за підручниками розв'язувати задачі з квантитативної лінгвістики, яку вивчають на третьому році навчання, але оптимізувати процес засвоєння інформації можна було б, якби існував застосунок, який генерував би ці задачі індивідуально для студента для практики в будь-якому місці, в будь-який час, і перевіряв правильність розрахунків.

Після цього був етап перевірки ідеї. Її затребуваність ми вже розуміємо, але є сенс проаналізувати, переконатися в її ефективності та перевагах для цільової аудиторії.

Наша ідея, насамперед, унікальна. Провівши попередній аналіз, було визначено, що аналогічних застосунків немає. Ми вже маємо підручники, у яких описано різні статистичні методи, якими можуть користуватися лінгвісти. Ці підручники містять приклади задач та їх розв'язання, вони є у різних форматах,

як паперові, так і електронні копії та версії. Але через свої першочергові інформативні функції вони не мають багатьох інструментів, які може забезпечити застосунок, як-от автоматична перевірка відповідей, генерування нових варіаційних рядів для тренування розв'язання задач стільки разів, скільки необхідно кожному студенту для якісного засвоєння вивченого матеріалу на індивідуальному рівні. Отже, створений продукт гарантує свою необхідність, хоч і для досить вузького кола споживачів.

Наступний етап – створення власне концепції продукту. Вона відповідає на два основні питання: що потрібно для успішного запуску нового продукту, та як це можна зробити. Концепція повинна включати основні характеристики, їхні цілі та завдання, переваги та особливості, і, звісно ж, цінність для користувачів. Саме з цього моменту ідея та концепція перетворюються на щось об'ємне та зрозуміле.

Основні характеристики продукту:

- Електронний застосунок у вигляді вебсайту;
- Доступність з будь-якого пристрою та в будь-якому місці з підключенням до мережі інтернету;
- Наявність інформаційної довідки необхідної для розв'язання задач із квантитативної лінгвістики;
- Кореляція задач із завданнями курсу;
- Можливість багаторазової генерації вхідних кількісних даних для задач одного типу;
- Можливість перевірки правильної відповіді;
- Зручний у користуванні та сучасний інтерфейс вебсайту.

Беручи до уваги перечислені характеристики, можемо сформулювати цінність продукту. Вона має бути всього одна, на відміну від кількості переваг, які ми можемо запропонувати, і саме вона відповідає на головне запитання користувача: чому ми повинні обрати саме цей продукт? Наша цінність – зручна практика розв'язування великої кількості задач з квантитативної лінгвістики будь-де та будь-коли.

Пройшовши усі ці етапи роботи та представивши концепцію науковому керівнику роботи, ми перейшли до розробки інфологічної моделі застосунку, що систематизує такі етапи роботи:

- Створення списку задач;
- Створення інформаційної довідки для студентів;
- Написання програмного коду вебсайту;
- Створення інтерфейсу вебсайту.

Детальний процес розробки ми описали далі.

3.2. Автоматична реалізація формування типів задач

Перше, що нам варто зробити – створити список задач, які покриватимуть програму курсу «Квантитативна лінгвістика» і які увійдуть до комплексу. Оскільки для нас основним джерелом інформації під час проходження курсу був підручник Валентини Перебийніс «Статистичні методи для лінгвістів» [Перебийніс 2002], то й приклади задач та інформацію для теоретичної довідки ми брали з нього.

Усі задачі повністю відповідають основним статистичним характеристикам та поняттям, які вивчаються в межах дисципліни: абсолютна частота, відносна частота, середня частота, середня відносна частота, міра коливання середньої частоти, середнє квадратичне відхилення, стандартна похибка відхилення середньої частоти, коефіцієнт варіації, коефіцієнт стабільності, Хі квадрат, критерій Стьюдента. Відповідно по завершенню навчання студенти повинні знати, як обчислити кожне з них, на яку поведінку мовленнєвих явищ вказують ці статистичні характеристики і як її варто трактувати.

Отож, список задач, які увійшли до комплексу:

- Обчислити абсолютну частоту слова / частини мови у тексті.
- Обчислити середню частоту слова / частини мови у тексті.

- Обчислити відносну частоту слова / частини мови (у відсотках) у вибірці довжиною X слововживань із абсолютною частотою одиниці в тексті У.
- Обчислити середню відносну частоту слова / частини мови (у відсотках) у вибірці довжиною X слововживань.
- Обчислити середнє квадратичне відхилення частот.
- Обчислити міру коливання середньої частоти із ймовірністю 68,3% / 95,5% / 99,7%.
- Обчислити стандартну похибку відхилення середньої частоти.
- Обчислити коефіцієнт варіації частот.
- Обчислити коефіцієнт стабільності частот.
- Обчислити значення Хі квадрат.

Усі задачі було розділено на три рівні складності, залежно від того, коли студенти вивчають конкретні статистичні характеристики за програмі курсу, та кількості обчислювальних дій, необхідних для одержання відповіді. До першого рівня увійшли задачі 1-4, до другого – задачі 5-9, до третього – задача 10. Це було зроблено для того, щоб студенти розв'язували лише ті задачі, які відповідають їхньому рівню знань, і відчували прогрес у навчанні після переходу на новий рівень.

На цьому етапі також було прийнято рішення додати нову функцію до рівнів 1 та 2 – генерування випадкової задачі. Тобто програма випадковим чином обиратиме одну з умов задач на рівні, на якому перебуває користувач, і пропонує її до розв'язання. Таким чином, студент не знає, яка задача зі списку випаде наступною, а отже, має вивчити теоретичний матеріал усієї дисципліни.

Якщо говорити про загальний принцип автоматичного обчислення статистичних параметрів, які вивчаються у курсі "Квантитативна лінгвістика", він залишається однаковим із тим, який ми використовували у стилеметричному дослідженні, оскільки формули обчислення цих показників одні. Тож програма обчислює значення за математичними формулами, які були програмно реалізовані у стилеметричному дослідженні (див. Додаток 6). Текстові

лінгвостатистичні дані були взяті із частотних словників чотирьох вибірок: текстів медійного стилю, наукового, художнього та офіційно-ділового.

Отже, задачі комплексу будуються за одним із восьми частотних словників. Частотні словники для слів та частин мови ми побудували раніше, але для роботи з задачами ми вирішили не використовувати частотні словники словоформ через те, що повторюваність таких одиниць у тексті дуже низька. Відповідно, матеріал задач – частотні словники лексем та частин мови для текстів медійного, наукового, художнього та офіційно-ділового стилів. Частотний словник лексем ми побудували на основі частотного словника словоформ. Множині слів, з яких випадковим чином обирається одне для умови задачі, відповідає множина лексем із ЧС лексем, а множині частин мов, з яких для умови задачі також обирається одна – множина частин мов з відповідного ЧС. У словниках за ними уже закріплені конкретні частотні ряди.

Таким чином, застосунок генеруватиме десятки тисяч варіантів задач з квантитативної лінгвістики, дозволяючи студентам тренувати свої навички стільки часу і разів, скільки необхідно. Відповідно, для згенерованої задачі застосунок пропонує ряд абсолютних частот із укладених ЧС – 50 варіантів за 50-ма підвибірками з існуючого частотного словника. Така кількість вибірок здається нам оптимальною для обчислення характеристик та цілком готує студентів до проведення складніших практичних робіт за межами застосунку.

Важливою перевагою таких матеріалів є можливість порівнювати між собою поведінку одиниць у текстах різних функційних стилів. Це дозволяє поглибити знання та розуміння студентами різних мовленнєвих явищ. У перспективі це може стати цінним та зручним у використанні матеріалом для лінгвістичних досліджень української мови.

Є ще декілька ситуацій, де ми повинні запропонувати студенту додаткову інформацію для розв’язання задачі. Наприклад, для задач 3 та 4 нам потрібна загальна кількість словоформ у тексті. Довжина кожної вибірки складає 50 000 словоформ, тож саме це число ми і вказуємо у цих задачах, воно залишається незмінним. Цікава також задача 6, у якій користувачеві необхідно обчислити

міру коливання середньої частоти чисел із тією чи іншою ймовірністю. За мірою коливання ми маємо лише три варіанти ймовірностей для визначення довірчого інтервалу: 68,3%, 95,5% і 99,7%, тому для задачі 6 програма випадковим чином підбирає одне з цих чисел та пропонує обчислити довірчий інтервал з відповідною точністю.

Розв’язавши задачу на папері або за допомогою калькулятора, студент може ввести свій варіант відповіді в окреме поле. Після цього програма порівнює власну відповідь із запропонованою відповіддю, і оголошує результат перевірки. Якщо користувач правильно запам’ятав та використав формулу певного показника, не допустив технічної помилки у розрахунках, застосунок повідомляє про правильність відповіді. Якщо відповідь користувача та програми не збігається, застосунок вказує правильний варіант, користувач має змогу перевірити власні розрахунки для знаходження помилки, а також розв’язати цю задачу ще раз, просто з іншим рядом частот. Весь процес запускається заново: генерування задачі та частотного ряду, обчислення за формулою, перевірка відповіді користувача.

Опишемо принцип генерування та перевірки кожної задачі.

Після натискання кнопки «Згенерувати задачу» JavaScript [JavaScript 2025] викликає PHP [PHP 2025] скрипт на сервері, який читає значення з бази та повертає дані на клієнтську сторону у форматі JSON [JSON 2025]. Після цього сторінка оновлює потрібні елементи отриманими новими значеннями та розраховує правильну відповідь. У випадку Рівнів 1 та 2 на сервер також передаються ті значення з випадних списків, які обрав користувач. Відповідно, обирається потрібен частотний словник, оскільки необхідно лише дві характеристики (стиль та одиниця) для того, щоб безпомилково ідентифікувати ЧС, із яким користувач хоче працювати. Після цього зчитується один випадковий запис у ЧС, який відповідає потрібним фільтрам. У Рівні 3 на сервер ніякої додаткової інформації не передається через відсутність випадних списків (цей рівень має лише одну задачу). Тому там перед SQL запитом випадково генеруються частина мови та два різні види текстів. При натисканні кнопки

«Перевірити» введена відповідь порівнюється із раніше розрахованою правильною із точністю 0,01. Після цього з’являється потрібне повідомлення.

3.3. Створення інтерфейсу навчального лінгвостатистичного застосунку

Для першого знайомства із сайтом дуже важливо, щоб він виглядав привабливо та був зручним у користуванні. Насамперед, користувачі дивляться, чи комфортно і приємно їм перебувати на сайті, і тільки потім вони починають знайомитися з ним детальніше. Інтерфейс – це візуалізація компонентів сайта чи додатка, з якими будуть безпосередньо взаємодіяти користувачі. Таким чином, при створенні інтерфейсу ми поставили перед собою два основні критерії: комфорт та простота, щоб студенти, які відвідали сторінку, мали позитивний досвід користування нею та із задоволенням поверталися ще раз.

Для розробки інтерфейсу застосунку (див. Додаток 7) ми використовували такі мови програмування:

- HTML [html 2025]: відповідає за “каркас” сторінки і визначає, де знаходиться кожен елемент: кнопки, заголовки, текст, поля для введення тощо;
- CSS [CSS 2025]: відповідає за зовнішній вигляд сайту, його дизайн: колір фону та кнопок, шрифти, додаткові візуальні елементи, адаптація для різних пристроїв;
- JavaScript [JavaScript 2025]: відповідає за інтерактивність застосунку та взаємодію з користувачем. Наприклад, саме частина програмного коду, написана JavaScript, відповідає за надсилання на сервер вибраних значень полів, заповнення полів новими значеннями, розрахунок правильної відповіді та перевірку відповіді користувача.

Також для створення застосунку використовувалася мова програмування PHP [PHP 2025], але з її допомогою виконуються усі потрібні запити до сервера, в тому числі генерація умов задач.

Отже, відкривши застосунок, користувач потрапляє на головну сторінку (див. Додаток 9), яка містить основну інформацію про призначення вебсайту та меню у верхній частині сторінки. Меню складається з таких пунктів:

- “Головна”: головна сторінка;
- “Інструкція”: вказівки з ефективного використання застосунку, пояснення навігації сайтом. Хоч навігація досить зрозуміла для пересічної людини, ми вирішили створити коротку інструкцію на випадок виникнення запитань щодо користування застосунком;
- “Рівень 1”, “Рівень 2” та “Рівень 3”: запропоновані задачі. Користувач може одразу вибрати рівень складності задач, які розв’язуватиме під час відвідування сторінки, особливо якщо це постійний користувач, ознайомлений із застосунком.

Внизу головної сторінки можна натиснути на кнопку “Далі”, яка посилає користувача на інструкцію з користування застосунком. Відповідна кнопка також є внизу інструкції, після натискання якої користувачеві пропонується обрати рівень задач, які він розв’язуватиме. Це зроблено для того, щоб зручно та послідовно представити комплекс вправ новому користувачеві, але ці кроки зовсім необов’язкові, тож постійні відвідувачі можуть відразу переходити до необхідного їм пункту меню.

Як ми писали раніше, секції “Рівень 1”, “Рівень 2” та “Рівень 3” містять задачі, запропоновані для розв’язання. Їхній список та розподіл можна знайти у параграфі 3.3 цієї роботи. З тієї причини, що рівні включають в себе по кілька задач, які базуються на восьми частотних словниках, ми пропонуємо обрати тип задачі, функційний стиль текстів і мовну одиницю (лексема чи частина мови). Натиснувши на відповідні поля, користувач бачить випадні списки та може обрати необхідну опцію із запропонованих. Наприклад, у випадному списку рівня 1 бачимо такі варіанти (див. Додаток 9):

- для функційних стилів: “Законодавчі тексти”, “Медійні тексти”, “Наукові тексти”, “Художні тексти”;
- для мовних одиниць: “Лексеми”, “Частини мови”;

- для типу задач: “Абсолютна частота”, “Середня частота”, “Відносна частота”, “Середня відносна частота”, “Випадкова задача”. Останній варіант пропонує для розв’язання випадкову задачу відповідного рівня.

Нижче бачимо кнопку “Згенерувати задачу”. При натисканні, очевидно, генерується умова задачі за заданим статистичним показником та додається частотний ряд, який відповідає рядку з частотного словника. Відразу під задачею знаходимо поле для варіанту відповіді користувача, у якому цифри можна вносити вручну, а також за допомогою регуляторів збільшити чи зменшити задане число. Інтервал регулятора становить 1. Єдиний виняток – задача 2 рівня 2 на обчислення міри коливання середньої частоти, де ми маємо два поля для відповідей, адже необхідно вести межі міри коливання для тієї чи іншої ймовірності (див. Додаток 9). Увівши свій варіант відповіді та впевнившись у його правильності, користувач повинен натиснути кнопку “Перевірити”, що знаходиться прямо під полем для введення числа. Якщо відповідь користувача збігається з варіантом застосунку, а отже, правильна, з’являється повідомлення: “Правильно! Добре пораховано”. Якщо відповідь неправильна, студент побачить повідомлення: “Неправильно. Правильна відповідь:”, яке пропонує правильний варіант відповіді.

На цьому етапі користувач має багато опцій: згенерувати нову задачу з цією самою умовою, обрати умови для іншої задачі у випадному списку або перейти на інший рівень.

Крім цього, наш застосунок адаптований для будь-якого пристрою, що означає зручність у користуванні незалежно від обставин: на занятті чи удома з використанням ПК, ноутбука чи планшета, в дорозі з телефона чи смарт-годинника. Навігація сайтом не змінюється залежно від пристрою, і забезпечує постійним доступ до застосунку, допоки є інтернет-покриття (див. Додаток 9).

Отже, ми змогли виконати як критерій комфорту, так і критерій простоти інтерфейсу нашого застосунку, зберігаючи якомога більше його функцій.

Завершений продукт ми назвали “Quantics”, взявши від назви квантитативної лінгвістики англійською мовою перші і останні літери: “**quant**itative linguis**tics**” (додаток 7).

Висновки до розділу 3.

У цьому розділі ми описали процес створення комплексу задач із квантитативної лінгвістики від ідеї до запуску. Ми визначили, що якісне навчання завжди потребує практичного компонента, і що цю функцію може виконувати створений продукт. Основна цінність застосунку – зручна практика розв’язування великої кількості задач з квантитативної лінгвістики будь-де та будь-коли. Після цього ми розбили розробку на кілька етапів, які мали привести нас до готового продукту. Ми написали список задач, які пропонує застосунок, та теоретичну довідку для повноти інформації, створили комфортний інтерфейс та інструкцію для навігації, втілили це у життя. Важливим етапом розробки було тестування розробниками та потенційними користувачами, яке дозволило звернути увагу на слабку сторони продукту та додати нові функції. Не зважаючи на це, застосунок має багато сторін для росту, які за потреби можна втілити в життя протягом наступних років.

ВИСНОВКИ

У межах кваліфікаційної роботи бакалавра було поставлено за мету укласти частотні словники для чотирьох стилів української мови: медійного, художнього, наукового та офіційно-ділового, за матеріалами ЧС провести стилеметричне дослідження і створити інтерактивний дидактичний застосунок для генерації задач з комп'ютерної лінгвістики. Мета досягнута, ми уклали вісім ЧС для чотирьох стилів, обчислили основні статистичні характеристики кожної вибірки, та створити навчальний застосунок (див. Додаток 7).

Для досягнення мети було виконано поставлені завдання, що дозволяє зробити висновки.

Квантитативна лінгвістика як наука базується на поєднанні лінгвістичних знань і математичних підходів до аналізу мови. Мову можливо досліджувати математично, адже це не лише система знаків, але й складна статистично зумовлена структура. Кількісні характеристики мовних одиниць, притаманні усім рівням мовної системи, дають змогу не тільки глибше зрозуміти мовні процеси, але й формалізувати лінгвістичні закономірності. В умовах інтенсивної комп'ютеризації науки з'являється необхідність автоматизовувати лінгвістичні завдання для глибшого дослідження мови та мовлення. З квантитативними методами у лінгвістиці активно працюють і сучасні українські мовознавці, але зібраних матеріалів для досліджень недостатньо, аби ґрунтовно дослідити кожен аспект української мови, тож такі дослідження варто продовжувати і заохочувати, а сам курс "Квантитативна лінгвістика" додавати в освітні програми філологічних дисциплін.

Було здійснено емпіричне дослідження чотирьох функційних стилів української мови — художнього, медійного, наукового та офіційно-ділового — шляхом обчислення статистичних характеристик текстів, репрезентативних для кожного з них. Усього було створено вісім частотних словників для кожної вибірки на 50 000 токенів: ЧС словоформ і ЧС частин мови. На цьому етап ми підтвердили, що мовлення дійсно надає перевагу невеликій кількості одиниць, адже в середньому лише 0,5% слів мали абсолютну частоту вище 100. Також ми

змогли побудувати діаграму розподілу відносних частот частин мов у текстах, яка показала значну різницю в частотах для різних частин мов. За матеріалами ЧС було проведено стилеметричне дослідження вибірок за базовими статистичними характеристиками та їхнім зіставленням. Згідно з результатами, найстабільнішим стилем виявилася художня проза, а найменш стабільним — офіційно-ділові тексти. Ці два стилі також найбільше відрізняються один від одного, що нам вдалося дослідити після зіставлення вибірок та обчислення критерію Стьюдента та ступеню розходження. Доведено, що статистичні розходження між вибірками є значущими, а отже — стилістичні відмінності між текстами належать до різних вибірок та не можуть об'єднуватися в одну. Одержані результати можуть бути корисними для класифікації текстів за стилями, характеристики цих стилів та інших мовознавчих досліджень.

У межах роботи було створено навчальний застосунок для автоматичного генерування і розв'язування задач із квантитативної лінгвістики. Застосунок надає студентам можливість доступно та зручно тренуватися у вирішенні задач будь-де й будь-коли. Створений застосунок став відповіддю на запит академічної спільноти і є унікальним ресурсом для вивчення квантитативної лінгвістики в українському освітньому просторі. Попри це, як і будь-який продукт, застосунок має свої можливості розвитку, які можна реалізувати в майбутньому, і в межах цієї роботи нам необхідно їх дослідити.

Є безліч методик проведення аналізу, і кожна має практичне застосування для тієї чи іншої проблеми. Одна із найбільш популярних і ефективних – SWOT-аналіз, тобто аналіз сильних та слабких сторін, можливостей та загроз. SWOT-аналіз використовується широким спектром організацій, від малого бізнесу до міжнародних корпорацій. Він застосовується в різних контекстах, включаючи розробку нових продуктів, вихід на нові ринки, а також управління змінами та ризиками в організації [Копчак 2024].

Через свою універсальність ми застосовували елементи цього методу для роботи з нашим продуктом. Під час створення ідеї та формулювання ціннісної пропозиції ми визначили сильні сторони, під час тестування виявили та по

можливості усунули недоліки та слабкі сторони. Настав час визначити потенційні сторони росту застосунку. Для цього ми провели аналіз функцій застосунку, які ми маємо зараз, врахували побажання та пропозиції бета-тестерів, які мали досвід користування продуктом, а також пригадали нереалізовані функції продукту, які ми не поставили в пріоритет поруч із реалізованою частиною.

1. Розширення списку задач. Для цього застосунку ми реалізували задачі, пов'язані з базовими статистичними характеристиками, але сама квантитативна лінгвістика не обмежується лише цими статистичними характеристиками або лише такими маніпуляціями з цими параметрами. Розширення списку задач, а також формату задач може спонукати студентів не лише до обчислення статистичних показників, але й до аналітичних роздумів, порівняння результатів та ґрунтовнішого розуміння статистичних мовленнєвих явищ.

2. Блок перевірки теоретичних знань. Робота з опрацюванням теорії квантитативної лінгвістики не була метою цієї роботи, але цілком може бути новою функцією застосунку. У такому випадку до кожної теми можна скласти список тестових запитань, які студенти проходитимуть для перевірки знань. Це допоможе зробити навчання більш інтерактивним та цікавим, сприятиме виявленню прогалин у знаннях та самонавчанню.

3. Можливість будувати інтервальні та варіаційні ряди. Цей тип задач ми не включили до поточної версії застосунку через складність реалізації ідеї, але це важливі теми для повного розуміння статистичних явищ у мовленні. Функція з побудови інтервальних і варіаційних рядів прямо на вебсайтіі, однозначно, стане цінною для засвоєння цієї інформації.

4. Вбудований калькулятор. Зараз єдиний випадок, коли користувач змушений перервати свою присутність на вебсайті – для використання калькулятора на пристрої. Потенційно ми маємо простір застосунку, який можна відвести для проведення розрахунків без відриву від сторінки задачі.

Цінність роботи полягає не лише у зібраних теоретичних узагальненнях та емпіричних результатах, але й у практичному інструменті, який можна активно

використовувати в навчальному процесі. Застосунок — це відповідь на реальні освітні потреби, він потенційно підвищить ефективність навчання, інтерес до нього, розширить доступ до інструментів лінгвістичного аналізу.

Науковий внесок роботи — це одержані результати аналізу функційних стилів української мови у вигляді баз даних, практичний внесок — створені частотні словники, які можуть бути джерелами для наступних стилістичних та статистичних досліджень, а також навчальний застосунок, який сприятиме розвитку квантитативної лінгвістики в Україні.

Довідкові матеріали:

1. CSS: CSS Documentation [Електронний ресурс]. Режим доступу: https://devdocs.io/css/
2. DB Browser: DB Browser for SQLite [Електронний ресурс]. Режим доступу: https://sqlitebrowser.org/
3. Excel: Microsoft Excel [Електронний ресурс]. Режим доступу: https://www.microsoft.com/uk-ua/microsoft-365/excel
4. html - HyperText Markup Language support [Електронний ресурс]. Режим доступу: https://docs.python.org/uk/3.8/library/html.html
5. JavaScript: JavaScript Documentation [Електронний ресурс]. Режим доступу: https://devdocs.io/javascript/
6. JSON: JavaScript Object Notation [Електронний ресурс]. Режим доступу: https://www.json.org/
7. PHP: Hypertext Preprocessor [Електронний ресурс]. Режим доступу: https://www.php.net/
8. pymorphy3 2.0.4 [Електронний ресурс]. Режим доступу: https://pypi.org/project/pymorphy3/
9. Python: Welcome to Python [Електронний ресурс]. Режим доступу: https://www.python.org/
10. re — Regular expression operations [Електронний ресурс]. Режим доступу: https://docs.python.org/uk/3.13/library/re.html
11. tokenize_uk 0.2.0 [Електронний ресурс]. Режим доступу: https://pypi.org/project/tokenize_uk/

Список використаних джерел:

1. Альтманн, Ґ., 2005. Мода та істина в лінгвістиці. *Проблеми квантитативної лінгвістики*, с. 3–11.
2. Безпояско, О.К., Городенська, К.Г. і Русанівський, В.М., 1993. *Граматика української мови. Морфологія: підручник*. Київ: Либідь.

3. Бук С., 2001. *Велика проза Івана Франка: електронний корпус, частотні словники та інші міждисциплінарні контексти : монографія.* Львів : ЛНУ імені Івана Франка.
4. Бук, С., 2008. *Основи статистичної лінгвістики: навчально-методичний посібник*. Львів: Вид. центр ЛНУ ім. Ів. Франка.
5. Бук, С., 2006. Статистичні характеристики лексики основних функціональних стилів української мови: спроба порівняння. *Лексикографічний бюлетень*, с. 166–170.
6. Вихованець, І., 2004. *Теоретична морфологія української мови: академічна граматика української мови*. Київ: Пульсари.
7. Відділ лексикології, лексикографії та структурно-математичної лінгвістики [Електронний ресурс]. Режим доступу: https://iul-nasu.org.ua/viddily-ta-naukovi-grupy/viddil-leksykologiyi-leksykografiyi-ta-strukturno-matematychnoyi-lingvistyky.html
8. Дарчук, Н. П., 2008. *Комп'ютерна лінгвістика (автоматичне опрацювання тексту)*. Київ: ВПЦ «Київ. ун-т».
9. Дарчук Н. П., 2005. Параметризована база даних сучасної української мови на основі частотних словників. *Проблеми квантитативної лінгвістики.* Чернівці: Рута, с. 100-110.
10. Зубань, О.М., 2020. *Комп'ютерне лексикографічне моделювання морфемної системи української мови: монографія*. Київ: ВПЦ «Київ. ун-т».
11. Зубань О. М., 2006. *Параметризована база даних як інструмент дослідження корпусу текстів.* Лексикографічний бюлетень: Зб. наук. пр. Київ: Ін-т української мови НАН України.
12. Капелюшний, А.О., 2007. *Практична стилістика української мови*. Львів: ПАІС.
13. Карпіловська, Є.А., 2006. *Вступ до прикладної лінгвістики: комп'ютерна лінгвістика*. Донецьк: Юго-Схід, Лтд.

14. Клименко Н.Ф., Карпіловська Є.А., Комарова Л.І. та ін., 2014. *Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система.* Клименко Н.Ф. Вибрані праці.
15. Копчак Ю. С., Лобунець Т. В., Луковський Р. І.,2024. *SWOT-аналіз як важливий інструмент у розробці стратегії бізнесу.* Економіка та суспільство. Випуск №61.
16. Кочерган, М.П., 2001. *Вступ до мовознавства.* ВЦ “Академія”.
17. Краснобаєва-Чорна, Ж.В., 2018. Квантитативні методи у лінгвістиці: новітні тенденції. *Науковий вісник ДДПУ імені І. Франка. Серія "Філологічні науки"*, 9, с. 93–97.
18. Кузнєцова Є. М., 2023. *Драбина.* Видавництво Старого Лева.
19. Кузнєцова Є. М., 2021. *Спитайте Мієчку*. Видавництво Старого Лева.
20. Кульчицький, І., 2021. Окремі аспекти квантитативних досліджень української мови. *Україна Модерна*, 27, с. 73–96.
21. Кучеренко, І.К., 2003. *Теоретичні питання граматики української мови: Морфологія*. 2-ге вид. Вінниця: Поділля-2000.
22. Лінгвостатистика. Спецкурс [Електронний ресурс]. Режим доступу: http://linguist.univ.kiev.ua/courses_stat.htm
23. Марков, А.А., 1913. Пример статистического исследования над текстом «Евгения Онегина». *Известия Академии наук по физико-математическому отделению*, 4, с. 153–162.
24. Мацько, Л.І., Сидоренко, О.М. і Мацько, О.М., 2003. *Стилістика української мови*. Київ: Вища школа.
25. Морозов Н.А., 1915. *Лингвистические спектры, как средство для отличения плагиатов от истинных произведений того или другого известного автора и для определения их эпохи*. Известия Академии Наук, т. XX.
26. Перебийніс, В.І., 2001. *Статистичні методи для лінгвістів: навчальний посібник*. Вінниця: Нова книга.

27. Перебийніс, В.С., Муравицька, М.П. і Дарчук, Н.П., 1985. *Частотні словники та їх використання*. Київ: Наукова думка.
28. Українська правда. [Електронний ресурс]. Режим доступу: https://www.pravda.com.ua/
29. Щербина, Ю.М., Шестакевич, Т.В. і Висоцька, В.А., 2010. Науковий напрям та навчальна дисципліна «Математична лінгвістика». *Інформаційні технології і засоби навчання*, 2, с. 34–42.
30. Ярошевич, І., 2012. Частини мови: поняттєво-термінологічний аспект. *Вісник НУ «Львівська політехніка». Серія «Проблеми української термінології»*, 733, с. 234–238.
31. De Morgan, A., 1847. *Formal Logic*. London: Taylor and Walton.
32. Greenberg, J.H., 1960. Quantitative Approach to the Morphological Typology of Language. In: *Universals of Language*. Cambridge, MA: MIT Press, с. 73–113.
33. Köhler, R. і Altmann, G., 2005. Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*, с. 12–41.
34. Lazaraton, A., 2000. Current Trends in Research Methodology and Statistics in Applied Linguistics. *TESOL Quarterly*, 34(1), с. 175–181.
35. mova.info: лінгвістичний портал [Електронний ресурс]. Режим доступу: http://www.mova.info/
36. Palmer, H.E., 1924. *A Grammar of Spoken English on a Strictly Phonetic Basis*. Cambridge: Heffer & Sons.
37. Swadesh, M., 1952. Lexicostatistical Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4), с. 452–463.
38. The Future of Jobs Report 2020 [Електронний ресурс]. Режим доступу: https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf
39. Zipf, G.K., 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

40.7 етапів розробки (NPD) New Product Development [Електронний ресурс]. Режим доступу: https://wezom.com.ua/ua/blog/etapy-razrabotki-new-product-development

ДОДАТКИ

Додаток 1. Диск з програмним кодом для укладання ЧС та файлами текстових вибірок:

https://drive.google.com/drive/folders/1JpBrZMg3YGXbbeQ8SOOgzZ1p4j8raCfp?usp=sharing

Програмний код:

https://drive.google.com/file/d/1iTqazIcKERDcG2sah9FzzzE7ReUgxux-/view?usp=drive_link

Вибірка медійних текстів:

https://drive.google.com/file/d/1FSEwQU24ikbPk8lk_BC0r7z3q6xZiZ_k/view?usp=drive_link

Вибірка художніх текстів: https://drive.google.com/file/d/16C-p4xRqg_BlUdzWGov9SaQPfz9J9fdA/view?usp=drive_link

Вибірка офіційно-ділових текстів:

https://drive.google.com/file/d/1Kve_7sO30hxSWv94o0x66JpNGsrXQuYb/view?usp=drive_link

Вибірка наукових текстів:

https://drive.google.com/file/d/1T5HuL2xmIGk3iP8QDtDW7X_fVtWJ_dGe/view?usp=drive_link

Додаток 2. Диск з повними базами даних:

https://drive.google.com/drive/folders/1GmgZu5dls8pfKA0bpECBYxleb81LnDAe?usp=sharing

Додаток 3. Фрагменти частотних словників.

Таблиця: Діловий_стиль_зведена

	word_form	lemma	part_of_speech	gen_freq
	Фільтр	Фільтр	Фільтр	Фільтр
1	та	та	CONJ	1201
2	в	в	PREP	763
3	до	до	NOUN	691
4	про	про	NOUN	652
5	у	у	PREP	646
6	за	за	ADV	632
7	з	з	PREP	593
8	спілки	спілка	NOUN	591
9	кредитної	кредитний	ADJ	573
10	на	на	INTJ	563
11	ради	рад	NOUN	490
12	наглядової	наглядовий	ADJ	457
13	України	Україна	NOUN	432
14	фінансової	фінансовий	ADJ	423
15	групи	група	NOUN	401
16	небанківської	небанківський	ADJ	365
17	що	що	CONJ	346

мал. 1. Слова з найбільшою абсолютною частотою (офіційно-діловий стиль)

Таблиця: Діловий_стиль_зведена

	word_form	lemma	part_of_speech	gen_freq
	Фільтр	Фільтр	Фільтр	Фільтр
5380	вирішувалися	вирішуватися	VERB	1
5381	зазначає	зазначати	VERB	1
5382	службами	служба	NOUN	1
5383	робота	робот	NOUN	1
5384	оцінюється	оцінюватися	VERB	1
5385	розроблених	розроблений	ADJ	1
5386	оприлюдненню	оприлюднення	NOUN	1
5387	профілю	профіль	NOUN	1
5388	представників	представник	NOUN	1
5389	професіоналізму	професіоналізм	NOUN	1
5390	компетентності	компетентність	NOUN	1
5391	різноманітності	різноманітність	NOUN	1
5392	прозорості	прозорість	NOUN	1
5393	принцип	принцип	NOUN	1
5394	посідають	посідати	VERB	1
5395	об'єктивного	об'єктивний	ADJ	1
5396	унітарного	унітарний	ADJ	1

мал. 2. Слова з найнижчою абсолютною частотою (офіційно-діловий стиль)

Таблиця: Художній_стиль_зведена

Filter in any column

	word_form	lemma	part_of_speech	gen_freq	sample_1	sample_2	sample_3	sample_4	sample_5	sample_6	sample_7	sample_8	sample_9	sample_10	sample_11	sample_12	sample_13	sample_14	sample_15	sample_16	sample_17
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	евгенія	евгеній	NOUN	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	кузнецова	кузнецов	NOUN	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	драбина	драбина	NOUN	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
4	багаж	багаж	NOUN	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	минулого	минулий	ADJ	8	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0
6	якби	якби	CONJ	34	2	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0
7	сюди	сюди	NPRO	19	1	0	0	1	0	0	0	0	1	0	0	0	0	2	1	1	0
8	хоч	хоч	PRCL	52	2	0	2	1	4	0	1	0	6	1	0	1	2	0	0	1	1
9	дві	два	NUMR	22	2	0	0	2	2	0	0	0	1	0	0	0	2	0	0	0	1
10	курки	курок	NOUN	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	сказала	сказати	VERB	254	5	0	3	1	7	1	4	3	4	0	0	5	5	7	1	4	1
12	притрівна	притрівна	NOUN	194	10	4	2	5	4	5	10	7	11	2	5	0	10	9	3	5	1
13	дивлячись	дивлячись	VERB	11	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14	через	через	PREP	54	1	0	0	0	1	1	1	1	1	0	1	2	2	2	1	2	0
15	кухонне	кухонний	ADJ	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	вікно	вікно	NOUN	15	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
17	надвір	надвір	ADV	16	1	0	0	0	0	0	2	0	2	1	0	0	0	1	1	0	0

мал. 3. Частотний словник слів (художній стиль)

Додаток 4. Програмний код для укладання ЧС та обчислення статистичних параметрів вибірок:

https://drive.google.com/file/d/1iTqazIcKERDcG2sah9FzzzE7ReUgxux-/view?usp=drive_link

Додаток 5. Фрагменти баз даних і результатів автоматичного обчислення статистичних характеристик вибірок.

Таблиця: Статистичні_даніАДЈ

Filter in any column

	x_i	n_i	xi_ni	x_average	xi_averagex	xi_averagex_2	xi_average_2_ni	sigma	sigma_average	v	v_max	d
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	111.0	1.0	111.0	191.7	-80.7	6512.49	6512.49	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
2	126.0	1.0	126.0	191.7	-65.7	4316.49	4316.49	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
3	135.0	1.0	135.0	191.7	-56.7	3214.89	3214.89	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
4	139.0	1.0	139.0	191.7	-52.7	2777.29	2777.29	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
5	140.0	1.0	140.0	191.7	-51.7	2672.89	2672.89	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
6	142.0	1.0	142.0	191.7	-49.7	2470.09	2470.09	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
7	143.0	1.0	143.0	191.7	-48.7	2371.69	2371.69	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
8	145.0	1.0	145.0	191.7	-46.7	2180.89	2180.89	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
9	148.0	1.0	148.0	191.7	-43.7	1909.69	1909.69	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
10	153.0	1.0	153.0	191.7	-38.7	1497.69	1497.69	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
11	154.0	2.0	308.0	191.7	-37.7	1421.29	2842.58	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
12	156.0	1.0	156.0	191.7	-35.7	1274.49	1274.49	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
13	158.0	1.0	158.0	191.7	-33.7	1135.69	1135.69	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
14	159.0	1.0	159.0	191.7	-32.7	1069.29	1069.29	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
15	161.0	1.0	161.0	191.7	-30.7	942.489999999999	942.489999999999	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
16	162.0	1.0	162.0	191.7	-29.7	882.089999999999	882.089999999999	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
17	166.0	1.0	166.0	191.7	-25.7	660.489999999999	660.489999999999	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782
18	170.0	1.0	170.0	191.7	-21.7	470.89	470.89	43.9569107194762	6.21644592995065	0.229300525401545	7	0.967242782

мал. 4. Статистичні характеристики прикметника (офіційно-діловий стиль)

Таблиця: Статистичні_даніNOUN

Filter in any column

	x_i	n_i	xi_ni	x_average	xi_averagex	xi_averagex_2	xi_average_2_ni	sigma	sigma_average	v	v_max	d
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	258.0	1.0	258.0	303.5	-45.5	2070.25	2070.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
2	273.0	1.0	273.0	303.5	-30.5	930.25	930.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
3	274.0	1.0	274.0	303.5	-29.5	870.25	870.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
4	278.0	2.0	556.0	303.5	-25.5	650.25	1300.5	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
5	280.0	1.0	280.0	303.5	-23.5	552.25	552.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
6	284.0	2.0	568.0	303.5	-19.5	380.25	760.5	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
7	289.0	3.0	867.0	303.5	-14.5	210.25	630.75	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
8	290.0	1.0	290.0	303.5	-13.5	182.25	182.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
9	291.0	3.0	873.0	303.5	-12.5	156.25	468.75	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
10	292.0	1.0	292.0	303.5	-11.5	132.25	132.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
11	293.0	1.0	293.0	303.5	-10.5	110.25	110.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
12	294.0	1.0	294.0	303.5	-9.5	90.25	90.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
13	295.0	1.0	295.0	303.5	-8.5	72.25	72.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
14	296.0	1.0	296.0	303.5	-7.5	56.25	56.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
15	297.0	2.0	594.0	303.5	-6.5	42.25	84.5	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
16	299.0	1.0	299.0	303.5	-4.5	20.25	20.25	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
17	301.0	3.0	903.0	303.5	-2.5	6.25	18.75	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715
18	302.0	2.0	604.0	303.5	-1.5	2.25	4.5	19.578815081613	2.7688625823612	0.06451009911569	7	0.9907842715

мал. 5. Статистичні характеристики іменника (художній стиль)

Таблиця: Статистичні_даніVERB

Filter in any column

	x_i	n_i	xi_ni	x_average	xi_averagex	xi_averagex_2	xi_average_2_ni	sigma	sigma_average	v	v_max	d
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	90.0	1.0	90.0	126.04	-36.04	1298.8816	1298.8816	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
2	95.0	1.0	95.0	126.04	-31.04	963.4816	963.4816	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
3	103.0	1.0	103.0	126.04	-23.04	530.8416	530.8416	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
4	105.0	2.0	210.0	126.04	-21.04	442.6816	885.363200000001	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
5	109.0	1.0	109.0	126.04	-17.04	290.3616	290.3616	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
6	111.0	1.0	111.0	126.04	-15.04	226.2016	226.2016	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
7	112.0	1.0	112.0	126.04	-14.04	197.1216	197.1216	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
8	113.0	2.0	226.0	126.04	-13.04	170.0416	340.0832	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
9	114.0	2.0	228.0	126.04	-12.04	144.9616	289.9232	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
10	120.0	2.0	240.0	126.04	-6.040000000000001	36.48160000000001	72.96320000000002	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
11	122.0	1.0	122.0	126.04	-4.040000000000001	16.3216	16.3216	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
12	123.0	1.0	123.0	126.04	-3.040000000000001	9.241600000000004	9.241600000000004	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
13	124.0	2.0	248.0	126.04	-2.040000000000001	4.161600000000003	8.323200000000005	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
14	125.0	1.0	125.0	126.04	-1.040000000000001	1.081600000000001	1.081600000000001	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
15	126.0	1.0	126.0	126.04	-0.040000000000063	0.001600000000005	0.001600000000005	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
16	127.0	5.0	635.0	126.04	0.9599999999999994	0.9215999999999988	4.607999999999994	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
17	128.0	2.0	256.0	126.04	1.959999999999999	3.841599999999998	7.683199999999995	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307
18	129.0	3.0	387.0	126.04	2.959999999999999	8.761599999999996	26.28479999999999	13.752032577041 1	1.9448310980648 2	0.1091084780787 14	7	0.98441307

мал. 6. Статистичні характеристики дієслова (медійний стиль)

Таблиця: Статистичні_дані4DV

Filter in any column

	x_i	n_i	xi_ni	x_average	xi_averagex	xi_averagex_2	xi_average_2_ni	sigma	sigma_average	v	v_max	d
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	12.0	1.0	12.0	33.02	-21.02	441.8404	441.8404	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
2	17.0	1.0	17.0	33.02	-16.02	256.6404	256.6404	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
3	19.0	3.0	57.0	33.02	-14.02	196.5604	589.6812	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
4	20.0	1.0	20.0	33.02	-13.02	169.5204	169.5204	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
5	22.0	1.0	22.0	33.02	-11.02	121.4404	121.4404	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
6	24.0	2.0	48.0	33.02	-9.02	81.360400000001	162.7208	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
7	25.0	4.0	100.0	33.02	-8.02	64.3204	257.2816	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
8	26.0	3.0	78.0	33.02	-7.02	49.2804	147.8412	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
9	27.0	1.0	27.0	33.02	-6.02	36.2404	36.2404	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
10	28.0	2.0	56.0	33.02	-5.02	25.2004	50.400800000001	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
11	29.0	3.0	87.0	33.02	-4.02	16.1604	48.481200000001	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
12	32.0	1.0	32.0	33.02	-1.02	1.0404000000001	1.0404000000001	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
13	33.0	3.0	99.0	33.02	-0.020000000000031	0.000400000000000125	0.001200000000000038	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
14	34.0	3.0	102.0	33.02	0.979999999999997	0.960399999999994	2.881199999999998	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
15	35.0	2.0	70.0	33.02	1.98	3.920399999999999	7.840799999999997	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
16	36.0	2.0	72.0	33.02	2.98	8.880399999999998	17.7608	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
17	37.0	1.0	37.0	33.02	3.98	15.8404	15.8404	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4
18	38.0	1.0	38.0	33.02	4.98	24.8004	24.8004	10.151827421701	1.4356852022640 6	0.3074480380681 3	7	0.9560793137 4

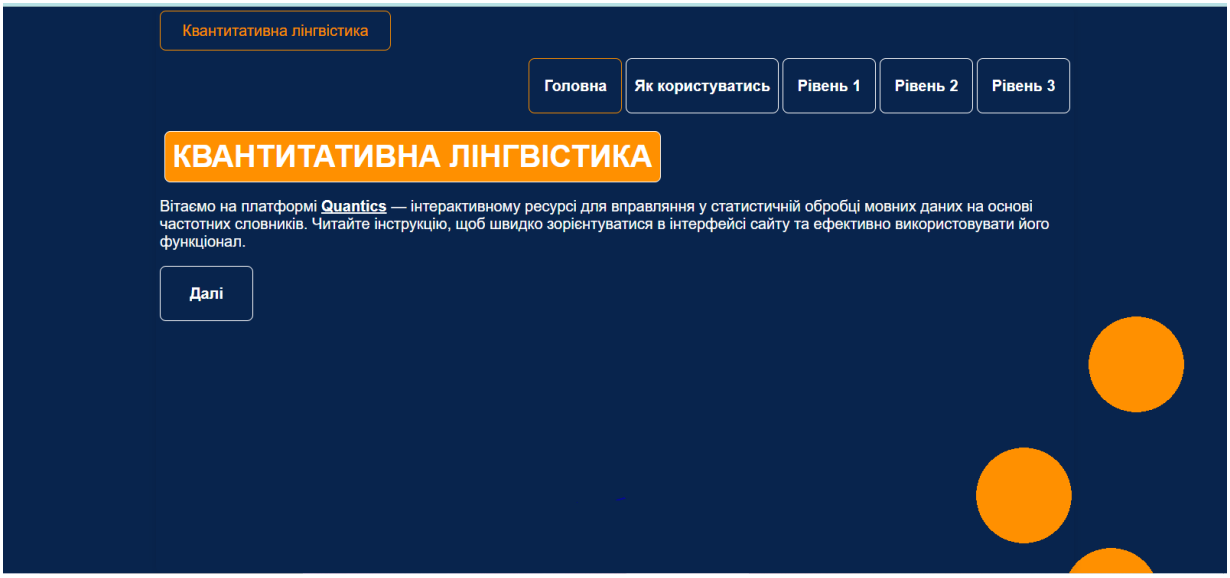
мал. 7. Статистичні характеристики прислівника(науковий стиль)

Додаток 6. Програмний код застосунку (архів):
https://drive.google.com/file/d/1Ts233KnRoSlYtGtr8Pxm2q-SBfqY3cne/view?usp=sharing

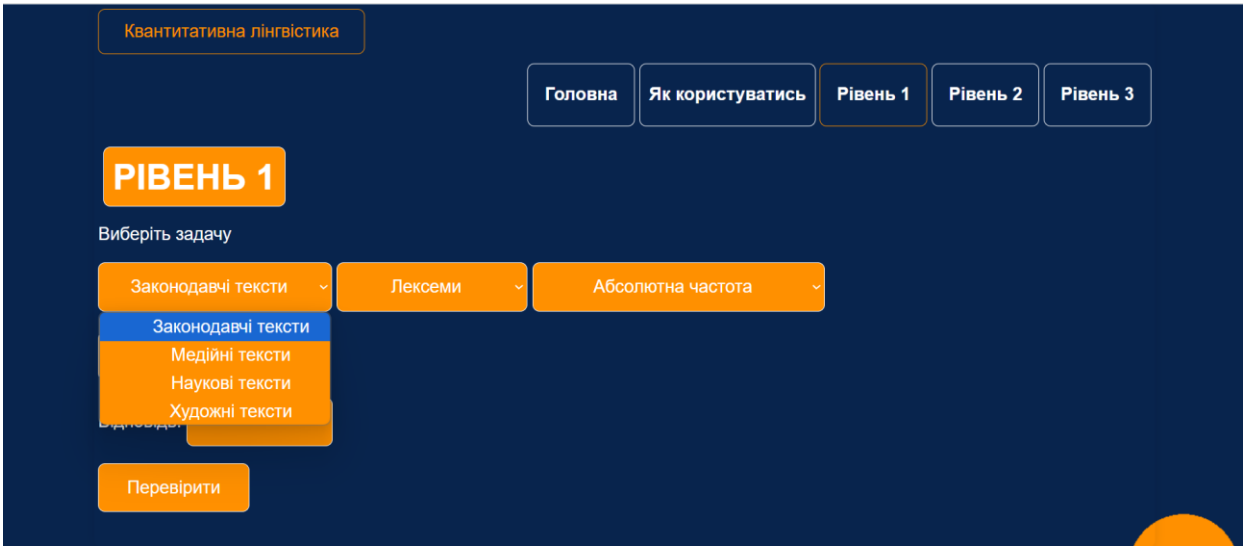
Додаток 7. Quantics. Режим доступу: https://quantics.fizma.net/index.php?idi=main/

Додаток 8. Відео роботи сайту:
https://drive.google.com/drive/u/4/folders/123CTrxjCnt9Ko7JUXjjr6G_t8uB7V8l_

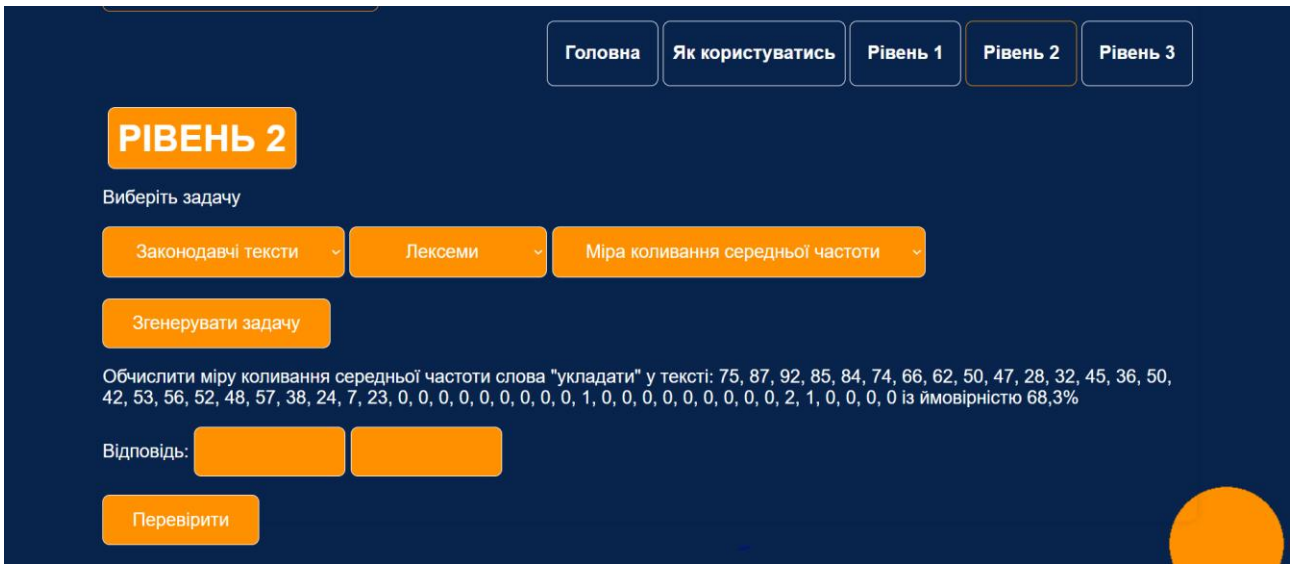
Додаток 9. Зображення елементів інтерфейсу застосунку Quantics.



мал. 8. Зображення головної сторінки платформи



мал. 9. Зображення з випадаючим списком (рівень 1)



мал. 10. Зображення задачі з обчислення міри коливання (рівень 2)

мал. 11. Адаптований інтерфейс на екрані смартфона