Дар'я Лебеденко, Аліна Трофімова-Герман
Національний технічний університет «Харківський політехнічний інститут»

ЕЛЕКТРОННИЙ ЗАСТОСУНОК ДЛЯ ІДЕНТИФІКАЦІЇ ТА ТЛУМАЧЕННЯ УКРАЇНСЬКИХ ДІАЛЕКТИЗМІВ У ТЕКСТАХ

У статті йдеться про розроблення програми, призначеної для ідентифікації та тлумачення українських діалектизмів у текстах. Оскільки українська мова багата на діалекти, проєкт має на меті впорядкувати процес виявлення діалектних особливостей у текстах, таким чином сприяючи лінгвістичним дослідженням і комунікації. Аналізуючи наявні словники українських діалектів і використовуючи методи оброблення природної мови, програма пропонує покращену доступність та ефективність розпізнавання діалектизмів.

Ключові слова: *українські діалектизми, розпізнавання діалектизмів, оброблення природної мови (NLP), корпус.*

This article deals with the development of a program designed to identify Ukrainian dialectic items in the texts. Since the Ukrainian language is rich in dialects, the project aims to streamline the process of identifying dialectic features in texts, thus contributing to linguistic research and communication. By analyzing existing dictionaries of Ukrainian dialects and using natural language processing methods, the program offers improved accessibility and efficiency of dialectism recognition.

Keywords: *Ukrainian dialectisms, dialectism recognition, natural language processing (NLP), corpus.*

З еволюцією суспільства розвивається і мова, що призводить до зростання потреби в інструментах, здатних аналізувати та

ідентифікувати мовні варіації, наприклад діалекти. Ця потреба виникає внаслідок впливу на мову мов сусідніх країн, політичних та територіальних особливостей, а також історичних та культурних факторів. Українська мова не стала винятком, поява різноманітних діалектів є історично обумовленою в ході її розвитку [1]. Саме тому актуальним є створення інструментів, що допомагають розпізнавати діалектизми в текстах, сприяючи розвитку мовних досліджень та комунікації. У цьому контексті мета цієї роботи — створення застосунку, що спрощує розуміння діалектизмів в україномовних текстах.

Для створення застосунку було проаналізовано словники українських діалектів та виявлено, що наявні словники існують лише у PDF-форматі або у вигляді сайту без можливості пошуку слів, що значно сповільнює з'ясування значень діалектизмів. Отже, розроблення застосунку для розпізнавання діалектизмів безпосередньо у тексті має такі переваги:

- підвищення доступності: програма надає користувачеві можливість швидко та легко знаходити діалекти в тексті;
- ефективність: застосунок швидко аналізує великі обсяги тексту, без необхідності використання будь-яких інших ресурсів, зокрема словників діалектів у PDF-форматі;
- покращення комунікації: знання і швидке розуміння діалектизмів допомагає покращити та спростити комунікацію між носіями різних регіональних варіантів мови;
- інтуїтивно зрозумілий інтерфейс: зручний і простий у використанні інтерфейс робить застосунок доступним для широкого кола користувачів, незалежно від їхнього віку або технічної підготовки;
- можливість подальших удосконалень: застосунок може бути легко оновлений і розширений новими функціями, такими як інтеграція з більшими лінгвістичними базами даних або адаптація для інших мов.

Розроблення застосунку для розпізнавання та аналізу діалектів у тексті здійснювалося в декілька етапів, а саме створення корпусу, дизайн інтерфейсу та реалізація алгоритму. Спочатку було зібрано діалекти та їх значення з надійних лінгвістичних джерел, таких як діалектні словники [3, 4, 5] та академічні публікації [2]. Під час оброблення створеного корпусу, що наразі налічує 6583 діалектизми, було використано методи оброблення природної мови (NLP), такі як розділення тексту на слова і фрази, створення множини унікальних діалектичних лексичних одиниць.

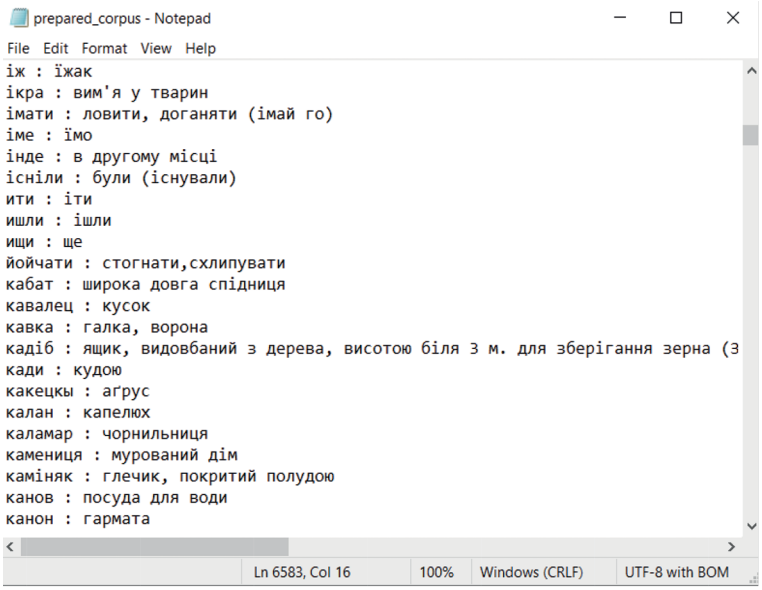

Рис. 1. Фрагмент створеного корпусу діалектів

Наступним кроком за допомогою HTML та CSS було створено макет вебсторінки програми 'DialectSpotter' з інтуїтивно зрозумілим дизайном. Сторінка складається з заголовка, простої форми введення тексту для аналізу та кнопки «Пошук». У разі знахо-

дження діалектизмів у тексті з'являється розділ під формою введення, у якому червоним кольором відображаються виявлені слова з відповідними дефініціями, поданими чорним кольором.

Для створення вебдодатка для аналізу діалектів, наявних у введеному користувачем тексті, використано фреймворк Flask. Також код імпортує бібліотеку pymorphy3 для морфологічного аналізу слів української мови. Створений додаток має два завдання: 1) відобразити головну сторінку з формою для введення тексту; 2) обробити введений текст і вивести знайдені в ньому діалекти з їхніми значеннями. Для досягнення мети було створено декілька функцій: read_dialects(), що читає дані про діалекти з файлу та повертає два списки: список діалектів і список їхніх визначень; to_infinitive(), що перетворює слова у їхню початкову форму, щоб полегшити порівняння; find_dialects(), що аналізує введений текст на наявність діалектизмів, використовуючи морфологічний аналіз pymorphy3.

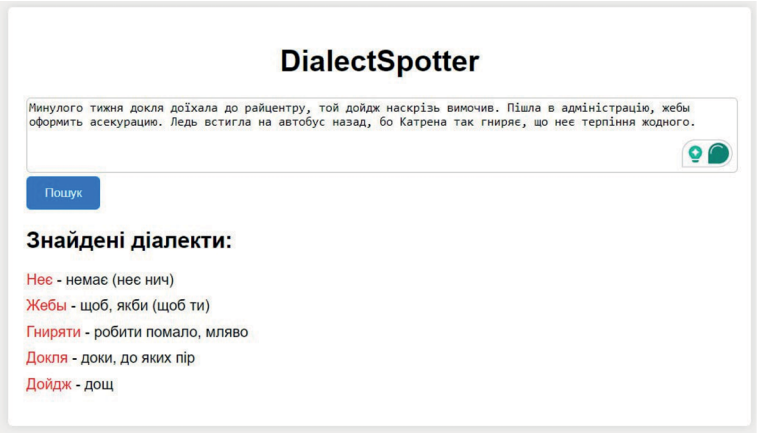

Рис. 2. Вигляд готового додатка

Застосунок «DialectSpotter» є зрозумілим у використанні навіть для недосвідченого користувача. Відкривши вебсторінку

застосунку, у спеціальному полі вводу користувач має ввести текст, який імовірно містить діалектизми. Після натискання кнопки «Пошук», введений текст надсилається на сервер для оброблення, де текст розбивається на окремі слова. Кожне слово перевіряється на відповідність діалектичним лексичним одиницям, що містяться у підготовленому корпусі. Застосунок використовує бібліотеку pymorphy3 для морфологічного аналізу слів, перетворюючи їх на нормальну форму. Для кожного слова у початковій формі застосунок перевіряє наявність у списку діалектизмів. Якщо знайдено збіг, діалектизм додається до списку знайдених слів разом із його тлумаченням. Список знайдених діалектизмів та їхніх значень виводиться на вебсторінці під полем вводу. Діалектизми виділяються червоним кольором для кращої наочності. Користувач бачить результати аналізу тексту без перезавантаження сторінки, що забезпечує зручність та швидкість використання застосунку. Такий процес повторюється кожного разу, як користувач вводить новий текст і натискає кнопку «Пошук».

В результаті дослідження розроблено та протестовано застосунок, що справно функціонує і наразі може розпізнати 6583 діалектизмів в україномовному тексті. Подальші перспективи розвиток програми можуть передбачати інтеграцію з більшими лінгвістичними базами даних, удосконалення алгоритмів розпізнавання діалектів та адаптацію для використання з іншими мовами. Таким чином, додаток 'DialectSpotter' є значним кроком до збереження та розуміння мовного різноманіття, притаманного українській мові.

Список використаних джерел

1. Діалекти української мови. Енциклопедія Сучасної України. URL: https://esu.com.ua/article-24451 (дата звернення: 03.05.2024).
2. Бігуняк А. В., Гойсак О. Г. Лемківський словничок. URL: http://www.lemko.org/lemko2/slovnyk.html (дата звернення: 01.05.2024).

3. Словник діалектних та маловживаних слів. Народна творчість. URL: https://proridne.org/Словник%20діалектних%20та%20маловживаних%20слів/#google_vignette (дата звернення: 13.05.2024).
4. Словник галицького діалекту. Інформаційно-аналітичний, культурно-просвітницький портал SPADOK.ORG.UA. URL: https://spadok.org.ua/etnolingvistyka/slovnyk-galytskogo-dialektu (дата звернення: 10.05.2024).
5. Москаленко А. А. Словник діалектизмів українських говірок Одеської області. 1958. LIBNAS | Library Portal of National Academy of Sciences of Ukraine. URL: http://irbisnbuv.gov.ua/cgibin/ua/elib.exe?Z21ID=&I21DBN=UKRLIB&P21DBN=UKRLIB&S21STN=1&S21REF=10&S21FMT=online_book&C21COM=S&S21CNR=20&S21P01=0&S21P02=0&S21P03=FF=&S21STR=ukr0001407 (дата звернення: 19.05.2024).