Таміла Крашган
Національний університет «Львівська політехніка»

СТВОРЕННЯ КОРИСТУВАЦЬКОГО ІНТЕРФЕЙСУ ДЛЯ ВЕСУМу

У цій роботі представлено нову пошукову систему, розроблену для Великого електронного словника української мови (ВЕСУМу). Метою цього проєкту є створення зручнішого інтерфейсу з ширшими можливостями пошуку для отримання більшої кількості інформації, що міститься у базі даних словника. Ця система дає змогу здійснювати пошук із використанням регулярних виразів та побудови пошукових граматик. Вона також надає доступ до структурованішого й зрозумілішого відображення інформації про знайдені леми.

***Ключові слова**: пошукова система, онлайн-словник, українська, ВЕСУМ.*

This paper presents a new search engine developed for the Large Electronic Dictionary of the Ukrainian Language (also known as VESUM). The aim of the current project is to set up a more user-friendly interface with broader search options, which at the same time provides more information contained in the Dictionary database. It enables the usage of wildcards in search queries and allows a user to set up search grammars. The developed system also provides a more structured and transparent way of displaying lemma information.

***Key words:** search engine, online dictionary, Ukrainian, VESUM.*

Великий електронний словник української мови (ВЕСУМ) створений у 2005 році як частиномовна база даних. Відтоді сам словник та його модифікації використовували в низці проєктів,

зокрема в пошукових системах української Вікіпедії та Генеральному регіонально анотованому корпусі української мови, а також у програмах перевірки орфографії української мови в LibreOffice чи LanguageTool [8]. Як наслідок, він містить не лише ті українські слова, які вважаються частиною літературної мови, а й спотворені, розмовні, діалектні та інші ненормативні форми, кожна з яких позначена відповідною позначкою.

Початковий варіант словника базувався на кількох друкованих словниках [1; 2; 6], а нині постійно поповнюється новими лемами, зокрема, немаркованими словами, знайденими в ГРАКу [3]. Сторінка словника на GitHub [13] містить останню версію бази даних і дозволяє користувачам вносити пропозиції виправлень і доповнень до словника.

Дані доступні у двох форматів: внутрішньому та вихідному. Внутрішній формат містить не усі словоформи, а лише самі леми та групи спеціальних тегів, що описують їх із граматичної та лексичної точки зору. З них формується вихідний формат бази — список лем та їхніх форм із меншими групами відповідних їм тегів. Його візуальне представлення і використовують у програмному забезпеченні.

Це саме те представлення, яке бачить користувач на сторінці ВЕСУМу при пошуку [7]. Ця сторінка надає лише базові опції пошуку: по лемах та словоформах без жодної можливості до використання пов'язаної з ними граматичної чи лексичної інформації. Це є дуже малою частиною того, що насправді можна було б зробити з усією інформацією, що доступна у базі даних. Звідси й виникає потреба створення нового користувацького інтерфейсу.

Перед розробленням нового інтерфейсу було проаналізовано наявні інструменти для близьких до української мов, а саме польської та білоруської. Граматичний словник польської мови [10; 12] надає різноманітні опції пошуку по лемах та їхніх формах: за лексичними класами, частотністю, родом тощо. Він

надає граматичну інформацію, таблиці відмінювання та — для деяких лем — роз'яснення значення. Недоліком цього словника є його неінтуїтивний інтерфейс, через який користувачам важко здійснювати розширений пошук без вивчення документації до словника. Граматична база білоруської мови [4; 5] також забезпечує інтерфейс пошуку, який використовує граматичну та лексичну інформацію лем і надає таблиці відмінювання. Порівняно з польським ресурсом, вона використовує більш візуальний і структурований спосіб фільтрації за параметрами лем. Крім того, вихідний код пошукової системи є загальнодоступним [11] і таким, що його можна налаштувати для інших мов.

На основі огляду онлайн-словників було зроблено висновок, що граматична база білоруської мови має найбільш відповідний пошуковий інтерфейс серед онлайн-словників слов'янських мов.

Однією з найважливіших відмінностей між ВЕСУМом і білоруською граматичною базою даних є формат, у якому зберігаються дані. У ВЕСУМі вони представлені у специфічних текстових форматах, тоді як база даних білоруської граматики використовує групу файлів XML, що описують парадигми, леми та відмінювані форми. Вона також містить групи тегів, що описують кожен з елементів. Однак граматична база сильно покладається на групи тегів, які мають певний жорсткий формат і порядок, тому використання запропонованого формату вимагало трансформації системи тегів ВЕСУМу. Для цього було створено додатковий допоміжний інструмент, який доступний на GitHub [9].

Із двох форматів представлення ВЕСУМу перетворення було здійснено для візуального, оскільки він надає набір даних, ближчий до опцій пошуку, які можуть зацікавити користувачів під час використання системи. Теги, які наразі підтримуються, надають інформацію про частину мови та специфічні частиномовні характеристики для лем та їхніх відмінюваних форм.

Наприклад, для іменників ознаками є такі: істота чи неістота, загальна чи власна назва, абревіатура чи ні, рід, число, відмінок.

Після виконання цього завдання потрібно було опрацювати кілька інших моментів для запуску оновленої пошукової системи ВЕСУМу, а саме: 1) перенесення пошукового функціоналу Korpus'у на більш компактну програмну структуру, 2) адаптація пошуку параметри для відповідності трансформованій системі тегів ВЕСУМу, 3) адаптація таблиць відмінювання для відповідності наборам відмінюваних форм, згенерованих у ВЕСУМі, 4) коригування дизайну, щоб зробити сторінку пошуку більш відповідною наявній екосистемі комп'ютерних лінгвістичних інструментів для української мови. За всіма перерахованими завданнями та подальшими кроками розвитку можна стежити на сторінці пошуку по ВЕСУМу на GitHub [14].

Налаштована пошукова система для VESUM надає користувачам можливості для виконання пошуку за лемами або за всіма відмінюваними формами в словнику, використовуючи як точні запити, так і регулярні вирази. Результати відображаються у вигляді лем зі списками їх граматичних особливостей. Натиснувши на лему, користувач може переглянути її таблицю відмінювання. Найпотужнішою частиною цього інструменту є пошукові граматики, з допомогою яких можна фільтрувати леми за їхніми ознаками. Укладаючи граматику, користувач може вибрати певну частину мови та частиномовні особливості, які його цікавлять.

Новий пошуковий інтерфейс впроваджує більш зручний спосіб взаємодії з базою даних, а завдяки розширеним параметрам пошуку дає змогу повніше використовувати інформацію, доступну для кожного з її елементів. Використання пошукових граматик у поєднанні з регулярними виразами може бути корисним для дослідників, яким потрібно скласти списки слів тої чи іншої форми із певною граматичною ознакою. Крім того, чітко структуровані таблиці відмінювання роблять словник

зручним для нелінгвістів, які можуть шукати правильне написання певних слів або їхніх форм.

Список використаних джерел

1. Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики / Є. А. Карпіловська та ін. ; за ред. Є.А. Карпіловської. Київ, 2013. 416 с.
2. Великий тлумачний словник сучасної української мови (з дод. і допов.) / гол. ред. В. Т. Бусел. Київ : ВТФ «Перун», 2005. 1728 с.
3. Генеральний регіонально анотований корпус української мови (ГРАК) / Шведова М. та ін. Київ, Львів, Єна, 2017–2022. URL: uacorpus.org (дата звернення: 16.05.2023).
4. Граматычная база : вебсайт. URL: https://bnkorpus.info/grammar.be.html (дата звернення: 16.05.2023).
5. Кошчанка У., Булойчык А. Граматычная база беларускай мовы. Мінск : Тэхналогія, 2021.
6. Граматичний словник української літературної мови. Словозміна: Близько 140 000 слів / В. І. Критська та ін. ; за ред. Н. Ф. Клименко. Київ : Вид. Дім Дмитра Бураго, 2011. 760 с.
7. Рисін А., Старко В. Великий електронний словник української мови (ВЕСУМ). Вебверсія 6.0.1. 2005–2022. URL: https://r2u.org.ua/vesum/ (дата звернення: 16.05.2023).
8. Старко В., Рисін А. Великий електронний словник української мови (ВЕСУМ) як засіб NLP для української мови. *Галактика Слова.* 2020.
9. Dictionary Format Translator : вебсайт. URL: https://github.com/tamila-krashtan/dictionary-format-translator (дата звернення: 16.05.2023).
10. Kieraś W., Woliński M. “Grammatical Dictionary of Polish” — an online version. *Jezyk Polski*. 2017. № 97. S. 84–93.
11. Korpus: Corpus Linguistics Software : вебсайт. URL: https://github.com/alex73/Software-Korpus (дата звернення: 16.05.2023).
12. Słownik gramatyczny języka polskiego : вебсайт. URL: http://sgjp.pl/ (дата звернення: 16.05.2023).
13. Project to generate POS tag dictionary for Ukrainian language : вебсайт. URL: https://github.com/brown-uk/dict_uk (дата звернення: 16.05.2023).
14. VESUM search : вебсайт. URL: https://github.com/tamila-krashtan/vesum-search (дата звернення: 16.05.2023).