Альона Дорожинська
(м. Київ)

ЛЕКСИКОН ПОЛЬСЬКОЇ ТА УКРАЇНСЬКОЇ АКТИВНОЇ ФРАЗЕОЛОГІЇ: СТВОРЕННЯ ОНЛАЙН СЛОВНИКА ІЗ ДРУКОВАНОЇ КНИГИ

Доступ до друкованих словників у наш час, зазвичай, обмежений. Це зумовлено перш за все малими тиражами (наприклад, Лексикон виданий накладом 1000 примірників). Представлення цього словника в онлайн форматі значно розширить його аудиторію. Крім того, сучасне цифрове лексикографічне середовище сформувало вже вимоги до лінгвістичних компетенцій своїх користувачів. Тому інтерфейси цифрових словників находяться у фокусі розробників API.

У цій роботі ми продемонструємо підхід до створення веб-сайту із паперового словника «Лексикон польської та української активної фразеології» (http://leksykon.loc/) [1].

Під час дослідження робота проводилась за таким планом:

1. аналіз структури тексту паперової книги;
2. візуальна розмітка файлів (HTML);
3. створення сайту на платформі WordPress.

Паперова книга

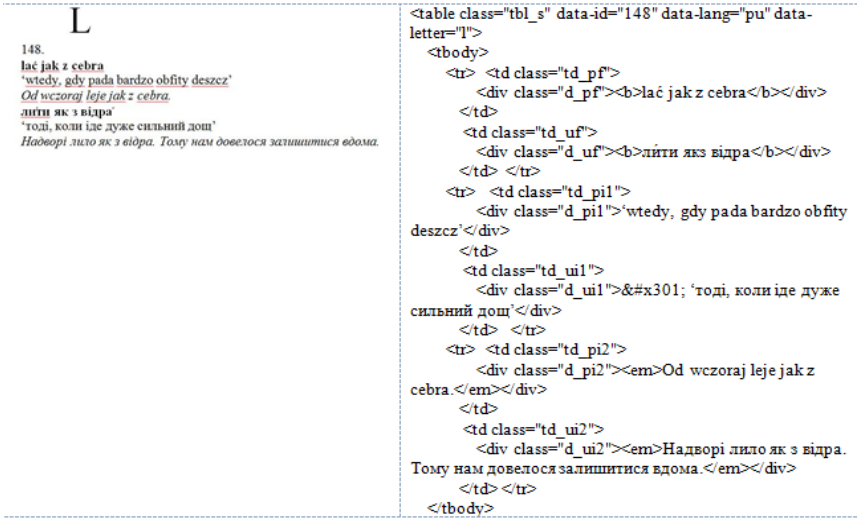
Перший крок нашої роботи – перетворенні файлу формату DOC у простий HTML-файл, що містить тільки візуальну розмітку. У нашому випадку нам пощастило отримати файли друкованого словника у форматі DOC: вступ, інформацію про авторів, список скорочень, корпус текстів словникових статей, та індекс. Всі файли ми отримали на польській і на українській мовах. Файли індексів ми на разі не використовували для створення статичного сайту [2].

Візуальна розмітка (HTML)

На другому етапі роботи ми виділили структурні елементи статей за їх формальними ознаками (шрифтовою розміткою в тексті) та розробили шаблон коду словникової статті. HTML файл нам необхідний для того, щоб правильно представити словникову статтю у редакторі матеріалів. В подальшому ця розмітка необхідна для побудови бази даних [3]. Як ідентифікатор структури словникової статті використовуються маркери HTML, які відповідно позначають:

< td class="td_pf"> − польський фразеологізм
< td class="td_uf"> − український фразеологізм
< td class="td_pi1"> − дефініція до польського фразеологізму
< td class="td_ui1"> − дефініція до українського фразеологізму
< td class="td_pi2"> − ілюстрація на польській мові
< td class="td_ui2"> − ілюстрація на українській мові

На рис. 1 показано приклад словникової статті у форматі DOC і візуально розмічену, з використанням представлених класів.



L

148.
lać jak z cebra
'wtedy, gdy pada bardzo obfity deszcz'
Od wczoraj leje jak z cebra.
лити як з відра
'тоді, коли іде дуже сильний дощ'
Надворі лило як з відра. Тому нам довелося залишитися вдома.

```
<table class="tbl_s" data-id="148" data-lang="pu" data-
letter="l">
  <tbody>
    <tr>  <td class="td_pf">
        <div class="d_pf"><b>lać jak z cebra</b></div>
      </td>
       <td class="td_uf">
        <div class="d_uf"><b>лити як з відра</b></div>
      </td> </tr>
    <tr>   <td class="td_pi1">
        <div class="d_pi1">'wtedy, gdy pada bardzo obfity
deszcz'</div>
      </td>
       <td class="td_ui1">
        <div class="d_ui1">&#x301; 'тоді, коли іде дуже
сильний дощ'</div>
      </td>  </tr>
    <tr>  <td class="td_pi2">
        <div class="d_pi2"><em>Od wczoraj leje jak z
cebra.</em></div>
      </td>
       <td class="td_ui2">
        <div class="d_ui2"><em>Надворі лило як з відра.
Тому нам довелося залишитися вдома.</em></div>
      </td> </tr>
  </tbody>
```

Рис. 1. Приклад представлення статті у форматах doc та html

Далі, для організації розмітки тексту відповідно до нашого шаблону, була написана програма, яка перетворює файл формату DOC у HTML файл. Таким методом ми перетворили весь корпус текстів словникових статей.

Інші файли було конвертовано в HTML-формат за допомогою редактора матеріалів WordPress.

WordPress

Для створення сайту ми використовували платформу WordPress — систему керування вмістом з відкритим кодом, яка через свою простоту в установленні та використанні широко застосовується для створення веб-сайтів. Сфера використання — від блогів до складних веб-сайтів. Вбудована система тем і плагінів у поєднанні з вдалою архітектурою дозволяє конструювати на основі WordPress практично будь-які веб-проекти.

Процес роботи із сайтом включає такі етапи:

1. Встановлення Денвера – локального сервера, на якому і буде розміщено WordPress.

2. Реєстрація в системі і початок створення сайту за допомогою тем, плагінів та шаблонів.

3. Розробка інтерфейсу сайту з використанням CSS каскадної таблиці стилів.

4. Розміщення алфавіту української та польської мов як окремої веб-сторінки.

5. Робота із вмістом сайту: наповнення сайту словниковими статтями польської та української фразеології, додавання вступу, інформації про авторів, умовних скорочень на обох мовах словника та створення вкладки *Контакти* для зворотного зв'язку. Візуальне представлення словникової статті на платформі WordPress можна побачити на рис 2.

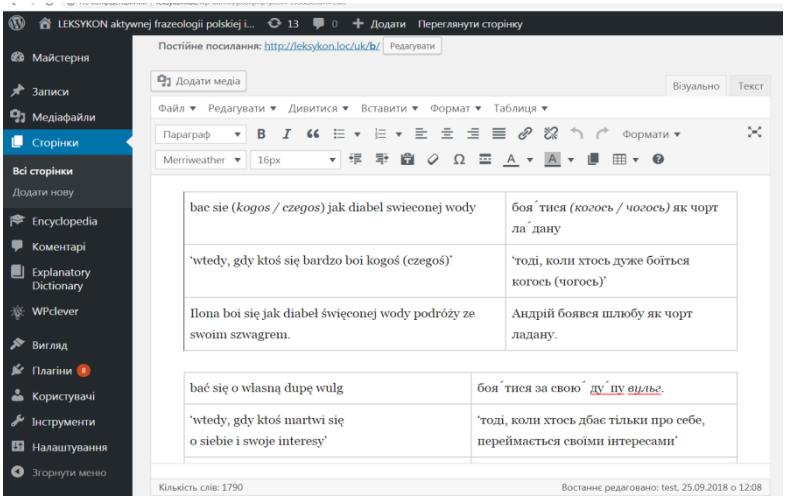
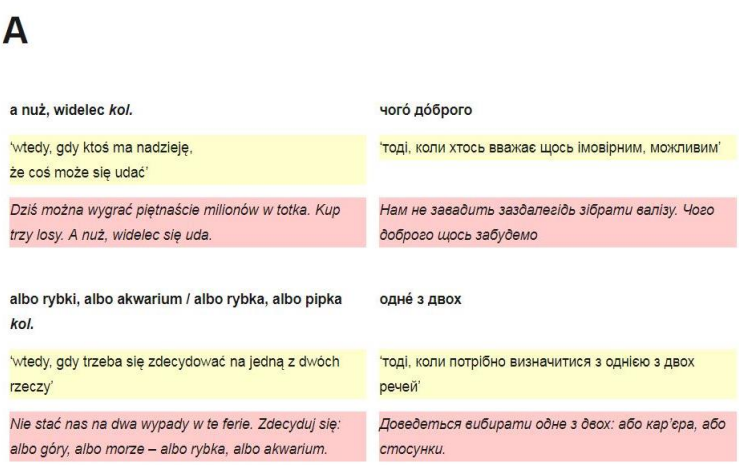


Рис. 2. Представлення статті у редакторі WordPress

6. Для того щоб представити виноски, було використано *Easy Footnotes* плагін, за допомогою якого створюються виноски у тексті. Вставляючи простий шорт код [note] [/note] отримаємо виноски.

7. Строго дотримані вимоги по структурі словникової статті, всі елементи розміщенні відповідно до її представлення у паперовій версії словника і підсвічені кольорами для кращої візуалізації (рис.3). У редакторі матеріалів було розроблено шаблон-контейнер для статті, щоб не порушити структуру і показати відповідність між елементами словникової статті.



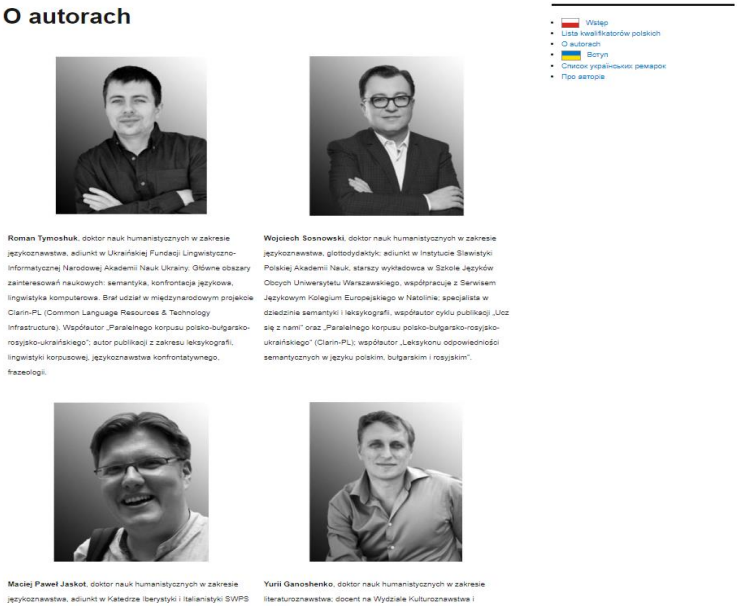
A

a nuż, widelec ***kol.***	**чого́ до́брого**
'wtedy, gdy ktoś ma nadzieję, że coś może się udać'	'тоді, коли хтось вважає щось імовірним, можливим'
Dziś można wygrać piętnaście milionów w totka. Kup trzy losy. A nuż, widelec się uda.	*Нам не завадить заздалегідь зібрати валізу. Чого доброго щось забудемо*
albo rybki, albo akwarium / albo rybka, albo pipka ***kol.***	**одне́ з двох**
'wtedy, gdy trzeba się zdecydować na jedną z dwóch rzeczy'	'тоді, коли потрібно визначитися з однією з двох речей'
Nie stać nas na dwa wypady w te ferie. Zdecyduj się: albo góry, albo morze – albo rybka, albo akwarium.	*Доведеться вибирати одне з двох: або кар'єра, або стосунки.*

Рис. 3. Фрагмент розміщених словникових статтей на сайті

8. Елементи меню справа створюються як окремі сторінки із вмістом (рис. 4)

Посилання на сайт на локальному сервері http://leksykon.loc/.



O autorach

- Wstęp
- Lista kwalifikatorów polskich
- O autorach
- Зміст
- Список українських ремарок
- Про авторів

Roman Tymoshuk, doktor nauk humanistycznych w zakresie językoznawstwa, adiunkt w Ukraińskiej Fundacji Lingwistyczno-Informatycznej Narodowej Akademii Nauk Ukrainy. Główne obszary zainteresowań naukowych: semantyka, konfrontacja językowa, lingwistyka komputerowa. Brał udział w międzynarodowym projekcie Clarin-PL (Common Language Resources & Technology Infrastructure). Współautor „Paralelnego korpusu polsko-bułgarsko-rosyjsko-ukraińskiego"; autor publikacji z zakresu leksykografii, lingwistyki korpusowej, językoznawstwa konfrontatywnego, frazeologii.

Wojciech Sosnowski, doktor nauk humanistycznych w zakresie językoznawstwa, glottodydaktyk; adiunkt w Instytucie Slawistyki Polskiej Akademii Nauk; starszy wykładowca w Szkole Języków Obcych Uniwersytetu Warszawskiego, współpracuje z Serwisem Językowym Kolegium Europejskiego w Natolinie; specjalista w dziedzinie semantyki i leksykografii, współautor cyklu publikacji „Ucz się z nami" oraz „Paralelnego korpusu polsko-bułgarsko-rosyjsko-ukraińskiego" (Clarin-PL); współautor „Leksykonu odpowiedniości semantycznych w języku polskim, bułgarskim i rosyjskim".

Maciej Paweł Jaskot, doktor nauk humanistycznych w zakresie językoznawstwa, adiunkt w Katedrze Iberystyki i Italianistyki SWPS

Yurii Ganoshenko, doktor nauk humanistycznych w zakresie literaturoznawstwa; docent na Wydziale Kulturoznawstwa i

Рис. 4. Сторінка, що містить інформацію про авторів

Головна сторінка сайту виглядає так, як на рис.5.

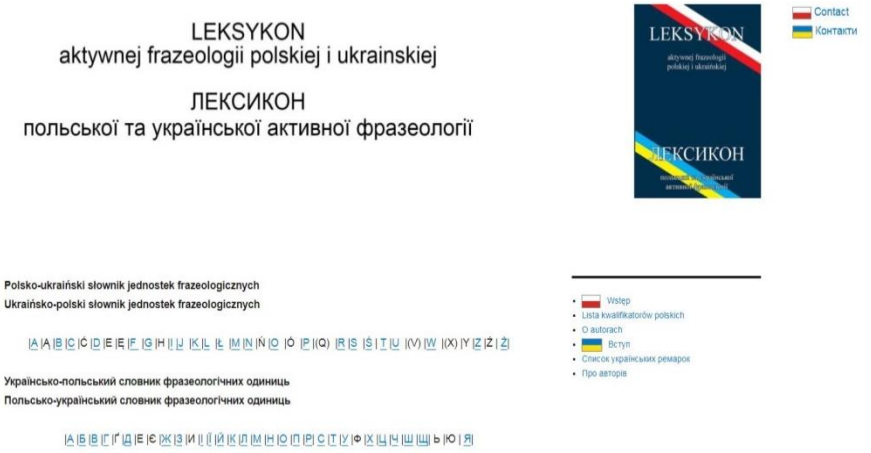


Рис. 5. Зовнішній вигляд сайту

Подальша робота над сайтом

Наступне завдання: організувати навігацію по статях Лексикону, використовуючи індекси словника (польський та український). Для цього необхідно представити структуру словника у форматі бази даних.

Нами була обрана швидкодіюча документо-орієнтована база даних LiteDB.

LiteDB не вимагає зовнішніх серверів бази даних і зберігає всі дані в переносимому файлі бази даних. Застосування бази даних такого типу надає можливість для прозорого представлення лексикографічних даних складної структури та побудови на її основі ефективних пошукових інтерфейсів.

Наступним кроком у роботі над проектом стане створення додатку для смартфонів,

У перспективі – створення на основі сайту контейнера для двомовних фразеологічних словників і ефективної технології конверсії відповідних текстів друкованих словників такого типу.

Література

1. Р. Тимошук, В. Сосновський, М. Яскот, Ю. Ганошенко. Лексикон польської та української активної фразеології. Варшава: JV DigitalSp.z o.o., 2018. 310 с.

2. Ефремова А. Н. Автоматизация распознавания макро- и микроструктуры бумажного словаря. *Слово и словарь=Vocabulum et vocabularium. Вып. 14*. Санкт-Петербург, 2016. С. 296−303.

3. Thomas Widmann, Phyllis Buchanan. The Orkney Dictionary: Creating an Online Dictionary Efficiently from a Printed Book. *Proceedings of the eLex conference*, 2017. С. 637−650.

Зубань Оксана, Гарбіч Аліна, Новікова Дар'я,
Романюк Богдана, Єсипенко Анастасія
(м. Київ)

БАЗИ ДАНИХ ЕЛЕКТРОННОЇ ВЕРСІЇ СЛОВНИКА «АКТИВНІ РЕСУРСИ СУЧАСНОЇ УКРАЇНСЬКОЇ НОМІНАЦІЇ» (студентський проект)

Сучасна українська лексикографія, відповідаючи вимогам інформаційного суспільства, характеризується тенденцією цифрового представлення лексикографічних даних. Керуючись актуальними завданнями й методами комп'ютерної лінгвістики у галузі словникарства, колектив студентів-бакалаврів 3-го курсу ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова» Інституту філології Київського національного університету ім. Т. Шевченка під керівництвом доцента кафедри української мови та прикладної лінгвістики О. Зубань розпочав проект «Електронна версія словника «Активні ресурси сучасної української номінації».

Мета проекту передбачає організацію роботи у два етапи: 1) автоматичну конвертацію текстової інформації у текстові дані: розроблення програмного забезпечення для автоматичного укладання баз даних[1] на матеріалі текстового файлу (формат *.txt) словника «Активні ресурси сучасної української номінації»[2] [1]; 2) конструювання електронної версії словника: розроблення програмних запитів до укладених баз даних і консольного інтерфейсу; розроблення людино-машинного інтерфейсу з використанням web-дизайну. Конференційна доповідь представляє результат перщого етапу роботи – структуру баз даних електронної версії Словника.

Словник «Активні ресурси сучасної української номінації» ознаменував своїм виходом у 2013 році новий етап української лексикографії – інтегральної неографії. «Мета Словника – описати нову українську лексику, що виявляє системотвірні ознаки, здатність до творення лексичних об'єднань: словотвірних гнізд і рядів, синонімічних рядів, антонімічних опозицій (пар або тріад), демонструє широкий спектр словосполук, а отже, виявляє розгалужені парадигматичні, синтагматичні й епідигматичні (деривaційні) відношення в тексті та в системі мови» [1, с. 8]. Інтегральний

[1] Бази даних позначаються у тексті статті абревіатурою БД.

[2] Словник «Активні ресурси сучасної української номінації» у тексті статті буде називатися скорочено Словник.