Standard Dependency Corpus for English. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).* 2014. P. 2897–2904. URL: https://aclweb.org/anthology/papers/L/L14/L14-1067/ (дата звернення 10.03.2019).

9. Universal Dependencies. URL: https://universaldependencies.org/ (дата звернення 10.03.2019).

Микита Яблочков
(м. Київ)

СТВОРЕННЯ КОМП'ЮТЕРНОГО ІНСТРУМЕНТАРІЮ ДЛЯ ДОСЛІДЖЕННЯ СТРУКТУРИ СТАТТІ ТЛУМАЧНОГО СЛОВНИКА: НА ПРИКЛАДІ СЛОВНИКА ІСПАНСЬКОЇ МОВИ

На основі аналізу інтерфейсів цифрових тлумачних словників (СУМ-20 – http:// http://services.ulif.org.ua/expl, Оксфордський словник англійської мови – http://en.oxforddictionaries.com/, Словник іспанської мови Королівської академії іспанської мови – http://dle.rae.es), що відрізняються структурами і методами укладання, виникла ідея створити інструмент для більш глибокого аналізу їх структур в одній концептуальній схемі. В цій статті описані перші етапи цієї роботи.

Цей інструментарій створюється як для аналізу структури статей тлумачного словника, так і усього словника в цілому, перевірки прозорості лексикографічних даних словника (виявлення імпліцитної інформації, структур та закономірностей), а також для можливості швидкої зміни зовнішніх інтерфейсів.

Як матеріал дослідження було вирішено використати текст Словника іспанської мови Королівської академії іспанської мови (http://dle.rae.es). Цей вибір був зумовлений наступним: принциповою відмінністю підходу до формування структури словника від прийнятої в УМІФ та тої, що використовується в Оксфордському словнику англійської мови (Oxford Dictionary), наявністю проведених досліджень лексикографічної системи іспанської мови [2], а також наявність цифрової версії словника у форматі HTML5, що гарантує автентичність тексту (його лінгвістичну достовірність) і дозволяє повністю зосередитися на його структурі.

Крім цього важливою особливістю цього словника є високий рівень складності структури словникових статей та велика кількість лінгвістичної інформації (включаючи граматичну, етимологічну та іншу), що включена в усі статті словника. Ця лексикографічна праця містить основну частину

національної лексики та фразеології й характеризується докладним описом лексико-граматичної і лексико-семантичної системи мови. Такий текст є носієм великого числа імпліцитно заданих зв'язків і відношень, що перетворює такі великі лексикографічні системи на певного роду «речі в собі».

Проведений аналіз структури словникових статей дозволив визначити первинні параметри для їх парсингу та подальшої обробки. Треба зазначити, що нас цікавили тільки параметри, що явно представлені в структурі словникових статей, а саме:

1. Омонімія:
   1.1. Присутня.
   1.2. Відсутня.
2. Опис реєстрової одиниці
   2.1. Заголовкове слово
   2.2. Член реєстрового ряду
   2.3. Дуплет (заголовкове слово)
   2.4. Дуплет (член реєстрового ряду)
3. Структура реєстрової одиниці
   3.1. Слово
   3.2. Словосполучення
   3.3. Морфема
4. Тип реєстрової одиниці
   4.1. Слово іншомовного походження
   4.2. Абревіатура
   4.3. Акронім
   4.4. Інше
5. Структура тлумачної частини статті
   5.1. Присутня дефініція реєстрового ряду
   5.2. Дефініція реєстрового ряду відсутня, присутня дефініція фразеологізму
   5.3. Дефініція реєстрового ряду відсутня, є посилання на неї
   5.4. Дефініція реєстрового ряду і посилання на неї відсутні, є посилання на дефініцію фразеологізму

Цей набір параметрів параметрів ми вважаємо достатнім для проведення начального аналізу словника.

Ефективний інструмент може бути побудован тільки при адекватном представлення складної структури словникової статті тлумачного словника в цифрових форматах. Накопичений досвід свідчить, що реляційні бази даних не дозволяють будувати достатньо ефективні цифрові лексикографічні системи. Зокрема, недоліком реляційних моделей при зберіганні та обслуговуванні лексикографічних об'єктів можна назвати той факт, що дані зберігаються імпліцитно – у вигляді набору з декількох таблиць та відношень

між ними. Оперування окремими таблицями як єдиним об'єктом вимагає потужної програмної інфраструктури і додаткових витрат часу для створення нових лексикографічних агентів, призначених для обробки цих масивів таблиць. Крім того, еволюційний потенціал цифрового об'єкту обмежений «непрозорістю» бази даних.

Оскільки словникові статті є елементами відповідної лексикографічної системи зі строго визначеною структурою, то логічним є їх пряме представлення у вигляді класів в об'єктно-орієнтованих мовах програмування, з подальшою обробкою, редагуванням та зберіганням саме в такому явному вигляді. Таку можливість дають NoSQL бази даних, а точніше – документно-орієнтовані бази даних, для яких головним елементом для зберігання і обробки є документ (об'єкт) зі строго описаною структурою.

Це дає можливість експліцитно зберігати лексикографічні об'єкти, не порушуючи при цьому їх внутрішню структуру, що відкриває прямий доступ до кожного елемента лексикографічного об'єкту і крім цього значно спрощує можливість їх редагування та модифікації (розширення).

Як приклад, у структурі словникової статті визначаються класи реєстрової та тлумачної частин, до них переходять атомарні елементи (реєстр та сама стаття відповідно) і з'являються нові елементи (характеристики реєстру та статті як цілісних об'єктів), після чого створюються класи для опису форми та змісту (характеристик) кожної одиниці реєстрового ряду (Схема 1).

При використанні реляційної бази даних така редукція лексикографічного об'єкту потребує крім створення додаткових таблиць ще й переписування функцій доступу та редагування, в той час як при використанні документарної бази даних новий складний об'єкт одразу зберігається в прямому вигляді. Інваріантність щодо даних, які зберігаються, – ще одна перевага документарних баз даних; єдиною вимогою є дотримання одноманітності структури об'єктів в одному сховищі (базі).

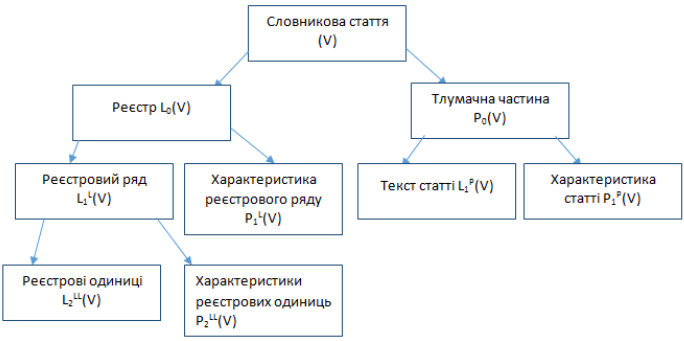

Схема 1

У процесі проектування було вирішено побудувати стенд для поступової модифікації додатку в бік розширення функціоналу (додавання нових лексикографічних агентів), а також поглиблення опису (редукції) лексикографічних об’єктів (словникових статей).

Первинний функціонал був реалізований без оптимізації роботи додатку та використання «важких» елементів інтерфейсу, через можливість змін в моделі лексикографічного об’єкту (статті) та супутніх змін в лексикографічних агентах, а також для демонстрації дослідницьких можливостей, що надають такий «легкий» інструментарій, використовуючи невелику кількість виділених характеристик.

Через ці ж причини було вирішено використати технологію Web API задля забезпечення обміну даними, а саме обробки запитів, між клієнтською та серверною частинами веб-додатку.

Для виконання закладених функції було створено наступні API: запит на отримання тексту статті за його ID та запит на отримання вибірки реєстра за параметрами.

Так як поставленою задачею був аналіз структури словника, інтерфейс розроблявся саме в такому напрямку – візуалізація статті і її структури, варіації пошуку по реєстру та створення вибірок словникових статей по раніше виявленим параметрам.

Основними елементами інтерфейсу являються вікно-список реєстру (Рис. 1, цифра 1), рядок з інформацією про кількість відображених реєстрових одиниць(Рис. 1, цифра 4)., вікно для відображення HTML-коду словникової статті (Рис. 1, цифра 3) та область відображення статті в оригінальному вигляді (Рис. 1, цифра 2).

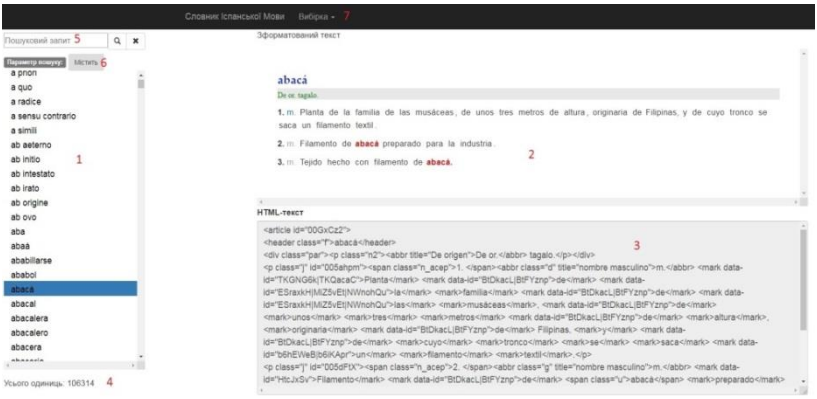

Рис. 1

Над списком реєстру знаходиться область з елементів інтерфейсу, що забезпечують декілька видів пошуку по реєстру(Рис. 1, цифра 5,6). Основним

елементом тут є віртуальна клавіатура для вводу пошукового запиту, що дозволяє набирати літери з діакритичними знаками іспанського алфавіту на пряму без встановлення іспанської мови на пристрій користувача (Рис. 2).

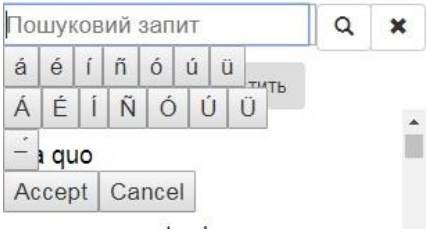

Рис. 2

Цей елемент можна легко модифікувати під відображення символів інших алфавітів шляхом модифікації його файлу налаштувань.

Зверху сторінки знаходиться кнопка виклику виринаючого вікна для вибору параметрів формування вибірки реєстру (Рис. 1, цифра 7).

Як приклад у цьому вікні можна вибрати характеристики «Заголовкове слово» з параметру «Структура реєстра», «Слово» з «Структура реєстрової одиниці», «Іншомовного походження» з «Тип реєстрової одиниці» та «Без омонімії» з «Омонімія» (див Рис. 3).

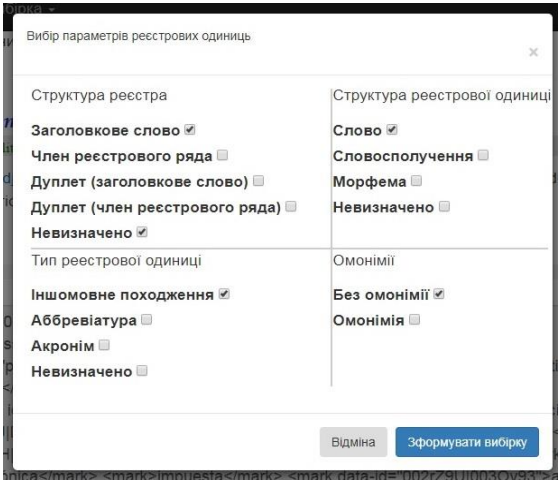

Рис. 3

Ця вибірка видає реєстрові одиниці, що являють собою заголовкове слово (не словосполучення) іншомовного походження, не маюче омонімів. Результат можливо побачити на Рис. 4.

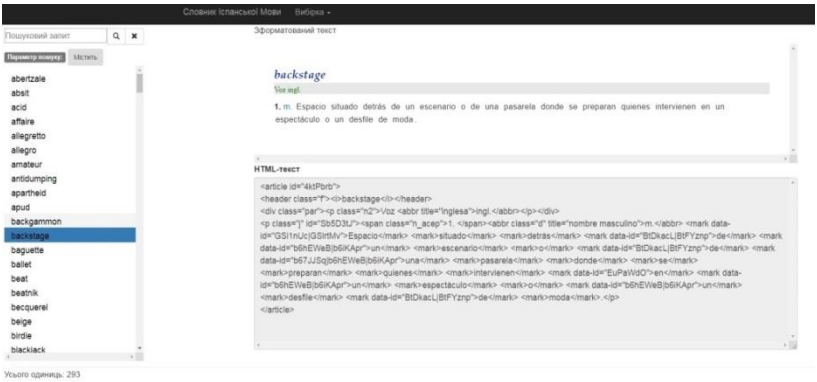

Рис. 4

Так як даний додаток планувався як стендовий, то його перша ітерація носила більш тестовий характер.

У подальшому планується здійснити редукцію класу тлумачної частини словникової статті за структурно-семантичними ознаками. Для цього буде здійснений аналіз статей з метою ідентифікації маркерів елементів структури словникових статей та розробка механізмів для їхнього індексування, пошуку та створення вибірок. Це дозволить виділити окремі семантичні тлумачення та більш глибоко розгорнути лексикографічну систему цього словника.

Також планується виділення граматичних і етимологічних ознак зі статей для подальшого створення на базі даного словника етимологічного і граматичного з подальшою їх інтеграцією.

У результаті планується отримати потужу інтегровану систему, спроможну підтримувати проведення різноманітних досліджень явних та прихованих лінгвістичних структур та елементів, що буде легко масштабуватися та модифікуватися під потреби різних словників.

Література

1. Bański Piotr, Bowers Jack, Erjavec Tomaž. TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms, Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. 2017. Pp. 485–494.

2. Купріянов Є. В. Лексикографічна система іспанської мови: феноменологія інтегрального опису: [монографія]. Київ: УМІФ НАНУ, 2018. 254 с.

3. Надутенко М. В. «Віртуалізовані лексикографічні системи та їх застосування у прикладній лінгвістиці»: автореф. дис. канд. тех. наук, НАН України, Нац. б-ка України ім. В.І. Вернадського. Київ, 2016. 22 с.

4. Широков В. А. Комп'ютерна лексикографія. Київ: Наукова думка, 2011. 351 с.