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.

характер Словника зумовив складність його макро- та мікроструктури: словникова стаття об'єднує 7 зон, які у свою чергу поділяються на підзони, а в деяких випадках підзони можуть структуруватися на менші підзони до 4-го рівня ієрархії; крім того, словникові статті, зони та підзони мають доповнювальні відношення у представленні лінгвістичної інформації макроструктури Словника. Складна лексикографічна модель Словника може бути значно спрощена для сприйняття та роботи користувача, якщо текстову інформацію в електронній лексикографічній системі представити окремими пошуковими зонами.

Конструювання електронної версії Словника базується на інфологічній та датологічній моделі баз даних, які передбачають структурацію на зони й підзони тексту Словника, відповідно до зонного принципу подання інформації в комп'ютерних версіях традиційних словників, розробленого Л. І. Колодяжною [4]. Текст словника уже має зонний принцип будови, тому інфологічна модель будується дуже просто: 1) на макрорівні вона відображає структуру Словника за розділами змісту; 2) на мікрорівні вона відображає ієрархічну залежність зон і підзон словникової статті. Особливістю даталогічного етапу є автоматична конвертація текстової інформації у дані: автоматичне конструювання БД електронної версії Словника.

За інфологічною моделлю макроструктури Словника, яка представлена у змісті (6 інформаційних зон словника), було створено даталогічну модель макроструктури, яка представляє 4 реляційні БД (див. Табл.1).

Таблиця 1. Кореляція зон інфологічної та даталогічної моделей

Інфологічна модель Словника	**Даталогічна модель електронної версії Словника**
1. Передмова	1. БД "база_передмови"
2. Умовні позначення у форматі статті	2. БД "умовні_позначення"
3. Умовні скорочення назв джерел	3. БД "умовні_скорочення"
4. Корпус словника	4. БД "корпусСловника"
5. Індекс слів	
6. Індекс словосполук	

Всі бази даних конструювалися автоматично за послідовністю виконання трьох завдань: 1) створення таблиці БД за допомогою системи керування базами даних SQLite3; 2) вилучення із тексту Словника (формат *.txt) за допомогою розробленого програмного забезпечення мовою Pyton текстової інформації і структурування цієї інформації на текстові частини: перетворення текстової інформації у текстові дані; 3) «пакування» таблиці БД: імпорт текстових даних у колонки таблиці.

Перші три БД (БД «база_передмови», БД «умовні_позначення», БД «умовні_скорочення») мають просту схему даних і кожна представлена однією таблицею, яка складається із трьох полів: 1) ID – номер об'єкта БД; 2) назва об'єкта БД; 3) текст, що відповідає пошуковому об'єкту.

БД «база_передмови» виконуватиме інформаційну функцію: знайомитиме користувача із лексикографічними особливостями Словника, його джерельною базою, будовою, призначенням.

БД «умовні_позначення» та БД «умовні_скорочення» будуть виконувати дві функції: 1) інформаційну – тлумачення значення умовного позначення та скорочення назви джерела; 2) інформаційно-навігаційну: кожне умовне позначення чи скорочення при зустрічі у тексті електронної версії Словника буде подано як гіперпосилання, до якого додаватиметься його пояснення.

Функціонально центральною в даталогічній моделі й найскладнішою за схемою даних є БД «корпусСловника». Вона складається із 9-ти таблиць, між якими встановлено зв'язок через поле ID – номер реєстрової одиниці, що в усіх таблицях однаковий (див. Рис.1).

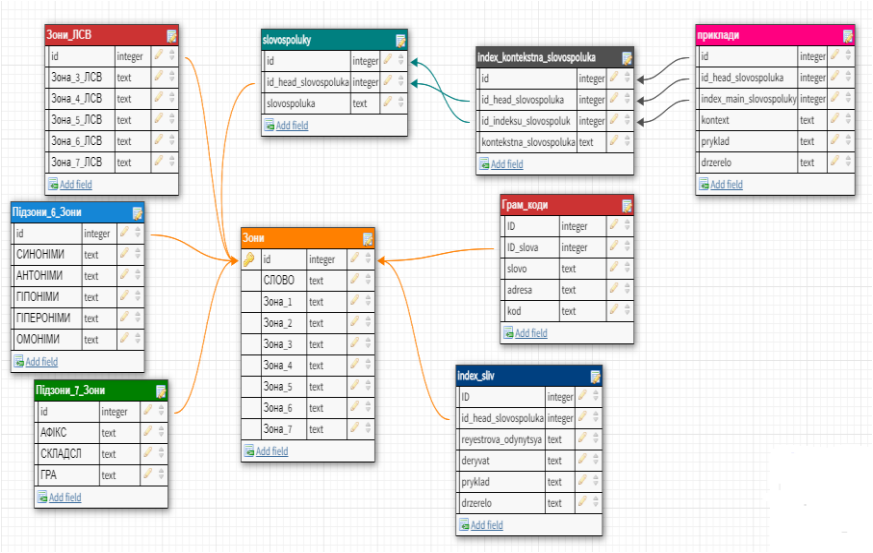

Рис.1. Реляційна діаграма схеми даних БД "корпусСловника"

БД «корпусСловника» виконує дві функції: 1) систематизує текстові дані статей реєстрових одиниць Словника [1, с. 22 – 360]; 2) забезпечує п'ять входів в електронний словник.

Систематизація текстових даних мікроструктури словника

Як демонструє реляційна діаграма на рис. 1, центральною у БД є таблиця «Зони», що систематизує текстові дані автоматичної сегментації тексту розділу «Корпус словника»: 1) на статті (статтею вважаємо текстову інформацію про одну одиницю реєстру (слово чи морфему) Словника від першої до сьомої зони включно); 2) у межах кожної статті на зони (зонами вважаємо пронумеровані у Словнику частини текстової інформації однієї статті від 1-ої до 7-ої).

Таблиці «Зони_ЛСВ», «Підзони_6_Зони», «Підзони_7_Зони» БД «корпусСловника» систематизують текстові дані про поділ 3-ої, 4-ої, 6-ої, 7-ої зон на підзони (підзонами вважаємо частини текстової інформації кожної зони статті, маркованої двома цифрами – 3.1., 3.2., 5.1., 5.2. і т. д.).

Деякі реєстрові одиниці Словника у межах зони 3 «Дефініція нова», а слово *зелений* і в зоні 4 «Дефініція стара» мають декілька ЛСВ, які марковані цифрами 3.1., 3.2., 3.3. або 4.1, 4.2. Важливо, що за цими ЛСВ закріплена відповідна лінгвістична інформація у 5-ій зоні – синтагматика, 6-ій зоні – парадигматика та 7-ій зоні – епідигматика. Для створення ефективного пошуку і класифікації інформації в електронній лексикографічній системі, було створено окрему таблицю «Зони_ЛСВ», в якій кожен ЛСВ реєстрової одиниці представлений окремим рядком із відповідною до них інформацією наступних 5-ої, 6-ої та 7-ої зон.

Кожна 6-та зона Словника «Парадигматичні відношення реєстрової одиниці (парадигм)» може складатися із 5-ти підзон: синоніми, антоніми, гіпоніми, гіпероніми та омоніми. З метою ефективного пошуку за кожним із 5-ти типів лінгвістичної інформації цієї зони текстова інформація 6-ої зони представлена у вигляді диференційованих текстових даних (окремі колонки) таблиці «Підзони_6_Зони».

Кожна 7-ма зона Словника «Епідигматичні (деривaційні) відношення реєстрової одиниці (епідигм)» може бути структурована на 3 підзони: 7.1. афікс – афіксальні прості похідні; 7.2. складсл – складні слова; 7.3. гра – мовна гра (каламбурне словотворення). Відповідно до цих підзон конструюється таблиця «Підзони_7_зони».

Таблиця «Грам_коди» укладається вручну за даними четвертої колонки «Зона_2» таблиці «Зони». У тексті Словника у 2-ій зоні подано граматичний код словозміни реєстрового слова (наприклад, 2. грам ж, -и, 1 О – код) за «Граматичним словником української літературної мови» [2], який представлено в електронній версій на порталі mova.info [3]. Із метою відображення моделі парадигми відмінювання реєстрового слова у створюваній електронній лексикографічній системі укладається таблиця «Грам_коди», яка поєднує граматичний код слова, взятий із 2-ї зони словника «Активні ресурси сучасної української номінації», та модель словозміни, що відповідає цьому коду в електронній версії «Граматичного словника

української літературної мови». Зв'язок граматичного коду із моделлю словозміни забезпечує 4-та колонка таблиці «adresa», що подає електронні адреси моделі словозміни електронної версії «Граматичного словника української літературної мови».

Таблиця «index_sliv» укладається з метою поєднання розділу «Індекс слів», що подає список дериватів, утворених на базі реєстрових одиниць, із контекстами вживання цих дериватів у 7-ій зоні словникової статті «Епідигматичні (деривaційні) відношення реєстрової одиниці (епідигм)».

Таблиці «slovospoluky», «index_kontextna_slovospoluka», систематизують текстові дані розділу «Індекс словосполук», у якому подано заголовкові слова, що входять до реєстру Словника, та словосполучення, які ці слова утворюють в ілюстративних текстах 5-ої зони словникової статті. Із цими таблицями пов'язана таблиця «приклади», яка зв'язує контексти вживання словосполучень, вилучені із 5-ої зони Словника, із реєстром словосполучень. Зв'язок між таблицями «приклади» та «index_kontextna_slovospoluka» здійснюється через 1-ше поле ID обох таблиць, в яких номери словосполучень збігаються. Це дозволяє представити однією константою (номером) різні тестові записи словосполучень: препароване словосполучення таблиці «index_kontextna_slovospoluka» та непрепароване словосполученя таблиці «приклади», наприклад: *альтернатив (у пошуках)* → ID 2=ID 2 ← *у пошуках альтернатив.*

Входи в електронний словник

Макроструктура текстового варіанта Словника має три входи до словникових статей: 1) за списком реєстрових одиниць розділу «Корпус словника» у змісті Словника [1, 3-5]; 2) за списком реєстрових одиниць та дериватів, утворених від реєстрових одиниць, у розділі «Індекс слів» [1, 360-388]; 3) за списком реєстрових одиниць та словосполучень із цими одиницями у розділі «Індекс словосполук» [1, 388-414]. В електронному варіанті Словника ставиться завдання збільшити кількість входів до п'яти. Функції п'ятьох входів в електронний словник виконуватимуть три центральні таблиці БД «корпусСловника»: 1) таблиця «Зони» – основний вхід (1) за списком реєстрових одиниць Словника (82 входи) до словникових статей; 2) таблиця «index_sliv» – додатковий вхід (2) за заголовковою реєстровою одиницею розділу «Індекс слів» до 7-ої зони Словника «Епідигматичні (деривaційні) відношення реєстрової одиниці (епідигм)»; додатковий вхід (3) за дериватами (2039 входів), утвореними від // за допомогою реєстрової одиниці, до прикладів текстового вживання 7-ої зони; 3) таблиця «index_kontextna_slovospoluka» – додатковий вхід (4) за заголовковою реєстровою одиницею розділу «Індекс словосполук» до 5-ої зони Словника «Синтагматичні відношення реєстрової одиниці (синтагм)»; додатковий вхід (5) за словосполученням (1868 входів) до прикладів текстового вживання 5-ої зони.

Велика кількість таблиць БД «корпусСловника» зумовлена автоматичним способом вилучення текстової інформації та імпорту текстових даних, що спричинило дублювання текстової інформації у деяких полях таблиць, тому на другому етапі конструювання електронної лексикографічної системи – розроблення програмних запитів та створення консольного інтерфейсу – кількість таблиць буде зменшена. Сьогодні колектив проекту працює над цим завданням. Найближчим часом електронна версія словника «Активні ресурси сучасної української номінації» буде представлена на порталі mova.info. Учасники проекту щиро дякують колективу авторів словника «Активні ресурси сучасної української номінації» за наданий електронний варіант тексту Словника.

Література

1. Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики / Є. А. Карпіловська, Л. П. Кислюк, Н. Ф. Клименко, В. І. Критська, Т. К. Пуздирєва, Ю. В. Романюк; Відп. ред. Є. А. Карпіловська. Київ: ТОВ «КММ», 2013. 416 с.

2. Граматичний словник української літературної мови. Словозміна / В. І. Критська, Т. І. Недозим, Л. В. Орлова, Т. К. Пуздирєва, Ю. В. Романюк; Відп. ред. Н. Ф. Клименко. Київ: Видавничий дім Дмитра Бураго, 2011. 760 с.

3. Граматичний словник української мови. URL: http://www.mova.info/grmasl.aspx.

4. Колодяжная Л. И. Структура словарного текста в аспекте машинной лексикографии: автореф. дис. ...канд. филол. наук. Москва, 1986. 23 с.

Наталія Клименко
(м. Покровськ)

ПРИНЦИПИ УКЛАДАННЯ СЛОВНИКА СХІДНОСТЕПОВИХ ГОВІРОК

Останнім часом активізувалося вивчення словникового складу новостворених українських діалектів. І чимала заслуга в цьому належить дослідникам слобожанських говірок. Ще на початку XXI ст. побачили світ такі лексикографічні праці, як: «Словник східнослобожанських українських говірок» (автори: К. Д. Глуховцева, В. В. Лєснова, І. О. Ніколаєнко, Т. П. Терновська, В. Д. Ужченко), «Словник діалектної лексики Луганщини» (відп. редактор – З. С. Сікорська). Приблизно в цей же час з'являється низка публікацій А. А. Сагаровського, присвячених укладанню діалектного словника Харківщини: («Лексика Центральної Слобожанщини (Харківщини) як об'єкт словництва» (2001), «Фрагмент діалектного словника Харківщини»