

*Остапова Ірина Вадимівна,
Широков Володимир Анатолійович
Український мовно-інформаційний фонд НАН України*

ЕТИМОЛОГІЯ В ЛАНДШАФТІ ЦИФРОВОЇ ЛЕКСИКОГРАФІЇ

Сьогодні цифрова лексикографія перетворилася на багатоаспектну галузь, яка є: а) прикладною науковою дисципліною на межі лінгвістики та інформатики, яка вивчає застосування методів і технологій інформатики до теорії та практики формування широкого спектру лексикографічних систем; б) галузь обчислювальної індустрії, яка швидко розвивається завдяки тому, що лексикографія знань є одним із ефективних способів їх ідентифікації та поширення. Застосування розробленої технології до етимологічних лексикографічних систем відкриває перспективи формування глобального лексикографічного середовища. При такому підході окремі словникові системи розглядаються як відносно самостійні об'єкти, і повнота лексикографічного етимологічного опису мов досягається завдяки їх інтеграції. Як інтегровальну структуру доцільно використати мовний індекс.

Ключові слова: *теорія лексикографічних систем, інформаційна система, етимологічний словник мови, віртуальна лексикографічна лабораторія, лексикографічне середовище*

Today the digital lexicographing transformed into a many-aspect branch which is: a) an applied scientific discipline at junction of linguistics and informatics, studying the ways and techniques to apply the methods of information science to theory and practice in forming a wide spectrum of lexicographic systems; b) a branch of computer industry which is developing rapidly mainly due to the fact that knowledge lexicographing is one of the efficient ways of its identification and propagation.

Application of the developed technology to the etymological lexicographic systems opens the prospects for formation of the lexicographic etymological medium. Under such approach the separate dictionary systems are considered as relatively independent objects and completeness of lexicographic etymological description of languages is reached due to their integration. As an integrating structure the language index may be used.

Keywords: *theory of lexicographic system, information system, etymological dictionary, virtual lexicographic lab, lexicographic medium*

Одним з важливих напрямів сучасної цифрової лексикографії є етимологія.

Етимологічний словник української мови в 7 томах [1] (далі – ЕСУМ) – фундаментальна лексикографічна праця. Перший том словника був надрукований в 1982, а шостий – в 2012 році. Сьомий том повинен являти собою багатомовний індекс до всього масиву словника. Текст словника включає понад 30 тисяч етимологій українських слів, при розробці яких було використано

273 мови. Для кожної з них необхідно створити окремий індекс з точними локалізаціями кожного слова – одиниці індексу.

Для мовного індексування словника було створено спеціальну комп'ютерну інструментальну систему – віртуальну лексикографічну лабораторію, яку адаптовано до структури ЕСУМ і орієнтовано на побудову індексів в автоматичному режимі. Цифрова реалізація ЕСУМ дала можливість доступу до всього корпусу словникового тексту незалежно від часу видання відповідного тому й відкрила можливості для різних цифрових інтерпретацій етимологічної інформації словника.

З метою створення ВЛЛ були виконані наступні кроки:

1. Побудова структури лексикографічної системи ЕСУМ відповідно до теорії лексикографічних систем [2];
2. Підготовка цифрового варіанту тексту ЕСУМ і виявлення елементів його метамови, які маркують структурні елементи лексикографічної системи;
3. Побудова структури бази даних, що відповідає структурі лексикографічної системи ЕСУМ з урахуванням маркерів його метамови;
4. Автоматична конверсія цифрового тексту ЕСУМ до лексикографічної бази даних.

Основна проблема створення ком'ютерних словників, виходячи з їхніх друкованих версій, це формування відповідної бази даних в автоматичному режимі безпосередньо з тексту словника (парсинг). Досвід переконує, що формування лексикографічних баз даних «вручну» для великих і складних словникових текстів практично неможливе. Основне завдання парсингу – автоматичне виокремлення структурних елементів словника безпосередньо з його тексту, оскільки саме вони виконують роль елементів лексикографічної бази даних.

Для ефективного вирішення завдання парсингу необхідна загальна теоретична схема опису і представлення об'єктів лексикографії. Як така використовується теорія лексикографічних систем [2].

Для виконання процедури автоматичного парсингу тексти всіх томів ЕСУМ були переведені у формат HTML й уніфіковані щодо структури файлів і знакової системи.

Застосування розробленої технології для інших етимологічних лексикографічних систем відкриває перспективу формування лексикографічного етимологічного середовища. При такому підході окремі словникові системи розглядаються як відносно незалежні об'єкти, а повнота лексикографічного етимологічного опису мов створюється за рахунок їхньої інтеграції. Як інтегровальну структуру доцільно використати мовний індекс.

Словник української мови [3] (СУМ) не містить етимологічної довідки. Отже, його інтеграція з ЕСУМ не лише дозволить збагатити лексикографічну структуру СУМ, але й надасть додаткову можливість простежити глоттохронологію.

Особливість Етимологічного словника української мови – фіксація для кожного реєстрового українського слова його еквівалентів у всіх слов'янських мовах. Для кожної слов'янської мови, відповідно до логіки системи, формується відповідний індекс. Це відкриває перспективи для інтеграції різних типів цифрових словників слов'янських мов.

Аналіз структур інших етимологічних словників доводить, що розроблена нами структура етимологічної лексикографічної системи, так само як і створений технологічний інструментарій, цілком ефективно можуть бути використані в інших цифрових етимологічних проектах, а також для інтеграції різних етимологічних ресурсів.

Література

1. Етимологічний словник української мови. Т. 1–6. Київ : Наукова думка., 1982–2012,
2. Широков В. А. Комп'ютерна лексикографія. Київ : Наукова думка., 2011. 356 с.
3. СЛОВНИК УКРАЇНСЬКОЇ МОВИ ONLINE. ТОМИ 1-10 (А-Ї). URL: <https://services.ulif.org.ua/expl/>. Title from the screen.