УДК 161.2.81”374.72”22
DOI https://doi.org/10.32841/2409-1154.2022.56.3

Бойко М. І.,
викладач кафедри української мови
Київського університету імені Бориса Грінченка

УКРАЇНСЬКА КОМП’ЮТЕРНА ЛЕКСИКОГРАФІЯ: СУСПІЛЬНІ ЗАПИТИ, ПРОБЛЕМИ ТА ПЕРСПЕКТИВИ

Анотація. Статтю присвячено аналізу стану української лексикографії на початку XXI століття. Актуальність визначається інтенсивним розвитком інформаційних технологій та відсутністю спеціального комплексного лінгвістичного дослідження, у якому не лише визначено теоретичні засади комп’ютерної лексикографії, а й здійснено огляд сучасних лексикографічних надбань. Мета статті – проаналізувати найпопулярніші електронні словники, визначити проблеми та перспективи української комп’ютерної лексикографії на початку XXI століття. Об’єктом дослідження є електронні словники на лінгвістичному порталі “Mova.info” та «Словники України» на Українському лінгвістичному порталі. Предмет дослідження – стан української комп’ютерної лексикографії на початку XXI століття. Подано огляд сучасних електронних словників. Становлення комп’ютерної лексикографії відбувається під великим впливом нових суспільних потреб і нових методів обробки інформації. З’ясовано, що ключовими напрямками діяльності є розробка та удосконалення інформаційно-лінгвістичних систем (традиційних і комп’ютерних словників, підручників, навчальних, експертних систем, систем автоматизованого опрацювання природної мови тощо, лінгвістичних баз даних та знань) та координація досліджень у галузі лінгвістичних технологій, комп’ютерної лінгвістики та лексикографії. Інтегрована лексикографічна система «Словники України» призначена для надання користувачам лексикографічної інформації про слова української мови. У системі «Словники України» на сьогодні представлено чотири словникові підсистеми-модулі – словозміна, синонімія, антонімія та фразеологія. Найпопулярніші онлайнові словники представлено на порталі “Mova.info”. Предметом нашого дослідження є «Електронний словник мови Тараса Шевченка», «Проєкт словника суржику», «Словник українських прийменникових колокацій» та інші. Варто акцентувати, що склад активних користувачів, способи, засоби, сама мова комунікації є надзвичайно динамічними – і тому потребують постійного детального вивчення. Очевидно, що сьогодні є потреба переведення лексикографічної спадщини людства у цифрову форму з метою нагромадження загальнокультурного потенціалу. Сучасні лексикографи повинні враховувати фундаментальні лексикографічні аспекти, а також навчитися визначати попит користувачів.

Ключові слова: комп’ютерна лексикографія, електронні словники, інформація, глобалізація.

Постановка проблеми. Проблема становлення української прикладної лінгвістики є однією з найважливіших і найскладніших в лінгвістичній науці. Варто наголосити, що на початку XXI століття комп’ютерна лексикографія постала як новий напрям людської діяльності, що, з одного боку, осмислив і втілив багатовіковий досвід теоретичних надбань лінгвістики та практичного лексикографування, а з іншого – отримав досі небачені перспективи.

Свого часу з цього приводу слушну думку висловив академік В. Широков, який зазначав, що словникова форма як тип суспільної комунікації все більше привертає увагу фахівців різних предметних сфер, бо за повнотою, змістовністю та точністю й правдивістю інформації словники посідають перше місце серед інших форм, які представлено в стислому вигляді [1, с. 3]. Згідно з концепцією вченого, словники не лише фіксують сукупність первинних знань, а й слугують надійним інструментом освоєння дійсності [1, с. 4].

Уважаємо, що активний розвиток української комп’ютерної лексикографії пов’язаний зі зростанням інформаційних технологій. Характерно, що проблема давно вийшла за межі теоретизування і має важливе прикладне значення, а тому нині активно обговорюється не тільки лінгвістами. Сьогодні відбувається неухильне зростання соціальної значущості лексикографічної продукції, тому скрупульозне дослідження надбань української комп’ютерної лексикографії є досить актуальною і важливою проблемою.

Аналіз останніх досліджень і публікацій. Проблему становлення та розвитку української комп’ютерної лексикографії вивчали В. Широков [1], О. Сидоренко [2], Є. Купріянов [3], Ю. Калимон [4] та інші вчені. Ґрунтовно проаналізовано сучасний стан вітчизняної комп’ютерної лексикографії та окреслено ключові проблеми галузі у монографії В. Широкова «Комп’ютерна лексикографія» [1]. Науковець не лише запропонував нову концепцію та комп’ютерну технологію лексикографічної діяльності, а й на основі ґрунтовного аналізу системотвірних відношень та інваріантів мовної субстанції здійснив огляд функціонування теорії лексикографічних систем на реальних прикладах створення об’єктів національної словникової бази. Аналізована теорія семантичних станів розширює уявлення про концептуальні технічні засади побудови багатомовних словникових систем та віртуальних систем професійної взаємодії в лінгвістиці.

Специфіку впровадження елементів комп’ютерної лексикографії в освітньому процесі розглянуто у розвідці О. Сидоренко «Українська комп’ютерна лексикографія як важливий інноваційний чинник навчального процесу» [2]. На думку дослідниці, варто створити єдиний лексикографічний комплекс та залучити до його наповнення щонайбільше фахівців з різних галузей знань [2, с. 528].

Є. Купріянов схарактеризував основні етапи розвитку комп’ютерних технологій укладання словників та визначив ключові чинники, що зумовили становлення комп’ютерної лексикографії в Україні [3]. Науковець впевнений, що бурхливий розвиток напряму пов’язаний з процесами інтеграції України

у світову інформаційну спільноту та створенням національної словникової бази.

У своєму дослідженні «Комп'ютерна лексикографія: виклики та перспективи» Ю. Калимон дослідила стан комп'ютерної (електронної) лексикографії на початку XXI століття, визначила характерні риси корпусного та електронного періодів. Ю. Калимон вважає, що розробка словників частково відбувається у віртуальній площині, у якій до обговорення реєстрових одиниць та інших аспектів укладання словника залучають безпосередньо і користувачів [4, с. 114].

На думку дослідниці, комп'ютерна лінгвістика використовує теоретичні та методичні досягнення інших наукових дисциплін (лінгвістики, лексикографії, комп'ютерних та корпусних технологій). Уважаємо, що актуальність вищезгаданої праці та новизна роботи полягає в аналізі сучасного стану досліджень, розробок та методик організації лексикографічної роботи [4, с. 116].

Однак, всупереч ґрунтовним напрацюванням, у цій царині мовознавства ще залишаються недослідженими окремі важливі сегменти. Актуальність визначається інтенсивним розвитком інформаційних технологій та відсутністю спеціального комплексного лінгвістичного дослідження, у якому не лише визначено теоретичні засади комп'ютерної лексикографії, а й здійснено огляд сучасних лексикографічних надбань.

Формування мети статті. Мета статті – проаналізувати найпопулярніші електронні словники, визначити проблеми та перспективи української комп'ютерної лексикографії на початку XXI століття. Реалізація мети передбачає розв'язання таких завдань: 1) з'ясувати особливості електронних словників (на прикладі словників, які представлено на лінгвістичному порталі "Mova.info" та Українському лінгвістичному порталі); 2) висвітлити шляхи розвитку української комп'ютерної лексикографії. Об'єктом дослідження є електронні словники на лінгвістичному порталі "Mova.info" та «Словники України» на Українському лінгвістичному порталі. Предмет дослідження – стан української комп'ютерної лексикографії на початку XXI століття.

Виклад основного матеріалу дослідження. Сьогодні людина не може уявити себе без комп'ютера, планшета, мобільного телефону, а особливо інтернету. Наприкінці XX – на початку XXI століття в усьому світі і в Україні засвідчено прискорення науково-технічного прогресу, становлення інформаційного суспільства, усеохопну глобалізацію й інтелектуалізацію переважної більшості сфер життя людей. Передусім цьому значною мірою сприяв інтенсивний розвиток інформаційних технологій. Більшість мовознавців впевнені, що на початку XXI століття лексикографія переживає особливий етап свого розвитку. Згідно з концепцією В. Широкова, становлення комп'ютерної лексикографії відбувається під великим впливом нових суспільних потреб і нових методів обробки інформації [1, с. 5].

Чималу кількість сучасних електронних словників підготовлено Українським мовно-інформаційним фондом Національної академії наук України. З'ясовано, що ключовими напрямками діяльності є розробка та удосконалення інформаційно-лінгвістичних систем (традиційних і комп'ютерних словників, підручників, навчальних, експертних систем, систем автоматизованого опрацювання природної мови тощо, лінгвістичних баз даних та знань) та координація досліджень у галузі лінгвістичних технологій, комп'ютерної лінгвістики та лексикографії.

Окрім того, фахівці Українського мовно-інформаційного фонду розробили інформаційно-лінгвістичну систему та створили лінгвістичний корпус, який має обсяг понад 200 млн слововживань і становить сучасних лексикографічних лінгвістичних досліджень.

Ґрунтовне вивчення специфіки Українського лінгвістичного порталу дає всі підставити стверджувати, що на сайті представлено перший в Україні повномасштабний електронний словник української мови – Інтегровану лексикографічну систему «Словники України». Електронне видання призначене для надання користувачам лексикографічної інформації про слова української мови.

У системі «Словники України» на сьогодні представлено чотири словникові підсистеми-модулі – словозміна, синонімія, антонімія та фразеологія. Виявлено, що базу генерального реєстру (близько 261 тис. слів) системи складає реєстр «Орфографічного словника української мови», 9-е видання (2009).

Вплив інтернету на повсякденне життя, а також на мову пересічних громадян є потужним і невпинно посилюється. Варто акцентувати, що склад активних користувачів, способи, засоби, сама мова комунікації є надзвичайно динамічними – і тому потребують постійного детального вивчення.

Ключове завдання лінгвістичного порталу "Mova.info: про українську мову, лінгвістику і не тільки" – мовне консультування суспільства. На сайті розміщено чимало наукових статей, монографій, підручників та інших матеріалів наукового характеру, пов'язаних з лінгвістикою, а також надано безоплатний доступ до інформаційних ресурсів (корпус української мови, електронні словники).

Досліджуючи вектори української комп'ютерної лексикографії, варто виокремити кілька лінгвістичних напрямів, які представлено на порталі "Mova.info":

- частотні словники;
- граматичний словник української мови;
- електронний словник мови Тараса Шевченка;
- проєкт словника суржику;
- відкритий словник;
- чотиромовний науковий словник;
- тезаурус з комп'ютерної лексикографії та тезаурус з лінгвістичної термінології;
- словник українських прийменникових колокацій та словник українських предикативних колокацій;
- семантичний словник української мови.

«Частотний словник сучасної української поетичної мови» репрезентує лексику поетичної мови, одержану на вибірці з поезій кращих українських митців кінця XX ст. обсягом 300 тис. слововживань. Будовою словника передбачено наведення статистичних даних як для парадигматичних форм, які зустрілися в текстах, так і для узагальнення словникових одиниць, що забезпечує можливість здійснення великого спектру граматичних досліджень. Наявний матеріал дає змогу констатувати, що запропонована методологія обробки лінгвістичних даних в електронній базі є узагальненням комплексу теоретичних і прикладних ідей сучасного мовознавства (рис. 1).

Проєкт «Відкритий словник виправлень суржику» створено з метою популяризації української мови серед населення. Словник є відкритою системою, оскільки усі зареєстровані

ПОЕТ: О. ЗАБУЖКО

ВСЬОГО ЗАПИСІВ: 9615

Слово	Частина мови	Абсолютна частота	Середня частота	Середньоквадратичне відхилення
артезіанським	ад'єктив	1	0,05	0,2179449
артеріях	ім.ж.р.	1	0,05	0,2179449
артисти	ім.ч.р.	1	0,05	0,2179449
архів	ім.ч.р.	1	0,05	0,2179449
архівах	ім.ч.р.	1	0,05	0,2179449
астенічних	ад'єктив	1	0,05	0,2179449
астронавтом	ім.ч.р.	1	0,05	0,2179449
асфальт	ім.ч.р.	1	0,05	0,2179449
асфальт	ім.ч.р.	1	0,05	0,2179449
асфальтам	ім.ч.р.	1	0,05	0,2179449

1 2 3 4 5 6 7 8 9 10 ...

Рис. 1. Результати пошуку в системі «Частотний словник сучасної української поетичної мови»

--	усе було так прозоро	усе було таке прозоре	Ева Гата роман „Очищення, або роблю, що хочу"	Львів, „БАК", 2007, стор. 85	непр.ф.	
--	рідина розподілялася по внутрішнім органам	рідина розподілялася по внутрішніХ органаХ	Володимир Горбатько, переклад з англійської Ден Браун „Втрачений символ", Харків, 2010	стор. 492	непр.ф.	
--	по деяким напрямкам	по деяких напрямах	Євген Нищук, міністр культури (!) ICTV, „Свобода слова"	20 лютня 2017, 22:30	непр.вжив.ф.	
--	подорож, якЕ змінило її життя	якА змінилА…	Валерій Калниш радіо „НВ", „Книга за книгою"	грудень 2018	непр.ф.	„путешествие" середнього роду, подорож – жіночого.

Рис. 2. Результати пошуку в системі «Відкритий словник виправлень суржику»

користувачі мають змогу створювати словникові статті, фіксувати помилкові форми з контекстом, зазначати правильний варіант відповіді та визначати категорію помилки (рис. 2).

«Словник українських прийменникових колокацій (усталених сполучень)» є результатом зіставлення частотних списків лем і словоформ усталених сполучень з підкорпусу законодавчих текстів Корпусу української мови. Реєстр словника нараховує понад 1415 колокацій 29 українських прийменників. За реєстрову одиницю прийнято усталене сполучення прийменника з певною формою колоката: відмінкові форми керованого іменника утворюють різні колокації, тоді як сполучення прийменника з колокатом у формі однини й множини належать до однієї колокації (керівник проекту – Т. Бобкова) (рис. 3).

Пошук за словоформою поза Знайти **Про словник**

Прийменник	Кількість словосполучень	Кількість слововживань
без	18	312
в	189	4998
від	48	934
для	101	2663
до	144	4686
з	268	6936
за	177	4920
з-за	1	42
зі	20	452
із	44	928
із-за	1	10
крім	7	114
між	8	128
на	211	6650
над	2	12
перед	2	30
під	14	1155
після	21	501
по	16	379
поза	1	13
при	33	646
про	135	3717
проти	1	14

Рис. 3. Результати пошуку в системі «Словник українських прийменникових колокацій (усталених сполучень)»

Також предметом нашого дослідження є «Електронний словник мови Тараса Шевченка», розроблений лабораторією комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка (керівник проекту – Н. Дарчук). Матеріалом для корпусу текстів письменника послужили його поетичні твори, написані українською мовою. Кожному зі слів у ході автоматичного морфологічного аналізу приписано морфологічні характеристики.

Варто акцентувати, що на основі цієї лінгвістичної інформації користувачі корпусу можуть здійснювати такі типи пошуку:

- певних лексем чи словоформ (задаючи в поле пошуку лише його початкові літери чи повну графічну форму);
- слів за морфологічними характеристиками (задаючи параметри частини мови, а для окремих частин мови вказуючи категорії: числа, відмінку, особи тощо);
- кореневих та афіксальних морфів (рис. 4).

Статистичні параметри

Лексем Всі

Показувати:
- ☑ Кількість текстів
- ☑ Середню частоту
- ☑ Середньоквадратичне відхилення
- ☐ Коефіцієнт стабільності

Показати

ВСЬОГО ЗАПИСІВ: 6889

Обмежити результат:
☐ За частотою з по

Слово	Частина мови	Абсолютна частота	Слово	Середня частота
веселіє	ім. с. р.	1	1	0,0163934426229508
веселля	ім. с. р.	1	1	0,0163934426229508
весело	прислівн.	15	12	0,245901639344262
весі	ім. множ.	1	1	0,0163934426229508
весілля	ім. с. р.	14	10	0,229508196721311
весільний	прикметник	1	1	0,0163934426229508
весна	ім. ж. р.	10	5	0,163934426229508
весною	прислівн.	7	5	0,114754098360656
весняний	прикметник	1	1	0,0163934426229508
веснянка	ім. ж. р.	1	1	0,0163934426229508
вести	дієсл.	15	14	0,245901639344262
весь	займ.-ім.	1	1	0,0163934426229508
весь	займ.-прикм.	108	65	1,77049180327869

Рис. 4. Результати пошуку в системі «Електронний словник мови Тараса Шевченка»

«Електронний словник мови Тараса Шевченка» може бути використаний для вивчання мови на високому рівні, також при підготовці навчальних матеріалів, вправ із використанням прикладів з реальних текстів та врахуванням частотної сполучуваності тощо.

Швидке й постійне оновлення лексики переконливо відбиває динаміку суспільно-політичного, соціально-економічного, наукового, технічного, побутового, ментально-когнітивного розвитку як української, так і світової спільноти. Згідно з концепцією Ю. Калимон, саме користувачі є рушійною силою для створення нових лексикографічних продуктів.

Висновки дослідження і перспективи подальших пошуків. Однак, обговорюючи становлення української комп'ютерної лексикографії, не можна не згадати важливі проблеми, які особливо актуалізувалися в наш час. Дехто з дослідників впевнені (В. Широков, Ю. Калимон та інші), що варто оновити фундаментальні лексикони, оскільки вони є зразками лінгвістичної думки певного історичного періоду.

Як зазначив В. Широков, сьогодні є потреба переведення лексикографічної спадщини людства у цифрову форму з метою нагромадження загальнокультурного потенціалу, а також використання в системах автоматичного опрацювання інформації [1].

На думку Ю. Калимон, з появою інтернету сучасна людина стала нетерплячою та прагне отримувати на будь-який свій запит швидку та конкретну відповідь, а отже концепція представлення знань у мобільних версіях словників має бути переглянута [4, с. 113].

Отже, сьогодні комп'ютерна лексикографія є багатоаспектною галуззю, прикладною науковою дисципліною на межі мовознавства та інформатики, яка вивчає способи та прийоми застосування методів інформаційної науки й технології у теорії та практиці створення широкого спектра лексикографічних систем; галуззю комп'ютерної індустрії, яка швидко розвивається головним чином завдяки тому, що лексикографування знань є одним з ефективних способів їхнього виявлення та поширення. Сучасні лексикографи повинні враховувати фундаментальні лексикографічні аспекти, а також навчитися визначати попит користувачів.

Уважаємо, що варто організовувати безоплатні лекції та презентувати суспільству надбання Українського мовно-інформаційного НАН України, адже цифрові лінгвістичні ресурси є невід'ємним складник якісної науково-дослідної діяльності. Перспективу дослідження вбачаємо в описі засад побудови електронних лексикографічних систем та ґрунтовному аналізі сучасних онлайнових словників.

Література:

1. Широков В. Комп'ютерна лексикографія : монографія. К. : Наук. думка, 2011. 351 с.
2. Сидоренко О. Українська комп'ютерна лексикографія як важливий інноваційний чинник навчального процесу. *Актуальні проблеми слов'янської філології*. 2010. Випуск XXIII. Частина 1. С. 524–528.
3. Купріянов Є. Комп'ютерна лексикографія як проблема сучасного мовознавства. 2008. URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/2480/1/2008_Kupriyanov_Kompiuterna%20leksykohrafiia.pdf
4. Калимон Ю. Комп'ютерна лексикографія: виклики та перспективи. *Актуальні питання іноземної філології.* 2019. № 10. С. 112–118.

Boiko M. Ukrainian computer lexicography: public requests, problems and perspectives

Summary. The article is devoted to the analysis of the state of Ukrainian lexicography at the beginning of the 21st century. The relevance is determined by the intensive development of information technologies and the lack of a special comprehensive linguistic study, which not only defines the theoretical foundations of computer lexicography, but also conducts an overview of modern lexicographic assets. The purpose of the article is to analyze the most popular electronic dictionaries, to determine the problems and prospects of Ukrainian computer lexicography at the beginning of the 21st century. The object of the research is electronic dictionaries on the linguistic portal "Mova.info" and "Dictionaries of Ukraine" on the Ukrainian linguistic portal. The subject of the study is the state of Ukrainian computer lexicography at the beginning of the 21st century. An overview of modern electronic dictionaries is presented. The formation of computer lexicography takes place under the great influence of new social needs and new methods of information processing. It was found that the key areas of activity are the development and improvement of information and linguistic systems (traditional and computer dictionaries, textbooks, educational and expert systems, automated natural language processing systems, etc., linguistic databases and knowledge) and coordination of research in the field of linguistic technologies, computer linguistics and lexicography. The integrated lexicographic system "Dictionaries of Ukraine" is designed to provide users with lexicographic information about words of the Ukrainian language. The "Dictionaries of Ukraine" system currently features four dictionary subsystems-modules – word change, synonymy, antonymy, and phraseology. The most popular online dictionaries are presented on the "Mova.info" portal. The subject of our research is the "Electronic Dictionary of the Language of Taras Shevchenko", "Surzhyk Dictionary Project", "Dictionary of Ukrainian Prepositional Collocations" and others. It is worth emphasizing that the composition of active users, methods, means, the very language of communication are extremely dynamic – and therefore require constant detailed study. It is obvious that today there is a need to transfer the lexicographic heritage of humanity into a digital form in order to accumulate the general cultural potential. Modern lexicographers must take into account fundamental lexicographical aspects, as well as learn to determine user demand.

Key words: computer lexicography, electronic dictionaries, information, globalization.