УДК 81'33

Т. В. Бобкова,
Київський національний лінгвістичний університет, м. Київ

УКРАЇНСЬКА КОРПУСНА ЛЕКСИКОГРАФІЯ: ОСНОВНІ ЕТАПИ Й ТЕНДЕНЦІЇ РОЗВИТКУ

У статті встановлюється періодизація української корпусної лексикографії, окреслюється типологія докорпусних і корпусних словників. Досліджується сучасний стан наявних у вільному доступі українських корпусів. Здійснюється спроба визначення типології сучасних корпусних словників за методологією укладання.

***Ключові слова**: корпусна лексикографія, корпусна лінгвістика, лексикографічний корпус текстів, докорпусний словник, корпусний словник.*

В статье устанавливается периодизация украинской корпусной лексикографии, определяется типология докорпусных и корпусных словарей. Анализируется современное состояние представленных в свободном доступе украинских корпусов. Предпринята попытка типологии современных корпусных словарей на основе методологии составления.

***Ключевые слова**: корпусная лексикография, корпусная лингвистика, лексикографический корпус текстов, докорпусный словарь, корпусный словарь.*

The article deals with Ukrainian corpus lexicography periodization and the main trends of pre-electronic, electronic corpora. The state of the art in Contemporary Ukrainian corpus linguistics is analyzed. The main trends of Ukrainian corpus lexicography is defined depending on the methodology.

***Keywords**: corpus lexicography, corpus linguistics, lexicographical corpus, pre-electronic corpus, electronic corpus.*

Актуальність дослідження основних тенденцій розвитку української корпусної лексикографії пов'язана з активним впровадженням корпусного підходу в сучасному мовознавстві. Це вимагає встановлення періодизації української корпусної лексикографії і критичного осмислення накопиченого досвіду, а також вирішення низки теоретико-методологічних питань щодо визначення лексикографічного корпусу текстів, окреслення типології сучасних українських корпусів текстів і корпусних словників.

Мета статті – дослідити етапи становлення, сучасний стан, основні тенденції розвитку української корпусної лексикографії з огляду на внесок у теорію й практику сучасного словникарства. Досягнення поставленої мети передбачає виконання таких **завдань**: 1) визначити історичні й концептуальні передумови становлення української корпусної лексикографії; 2) дослідити еволюцію типології українського корпусного словника; 3) встановити етапи й основні тенденції розвитку української корпусної лексикографії; 4) здійснити спробу типології сучасних корпусних словників на підставі методології укладання.

У сучасному мовознавстві корпусна лексикографія визначається як лінгвістична дисципліна, яка вивчає теорію і практику укладання корпусних словників. Корпусна лексикографія базується на понятті корпусу текстів, під яким розуміється електронний ресурс, що зазвичай містить величезну кількість слів з багатьох різних джерел [12, с. 270]. Отже, на відміну від традиційної лексикографії, лексикографічним або лексичним джерелом для укладання корпусного словника слугує електронний корпус текстів. Провідною ідеєю корпусного словникарства є, насамперед, твердження про можливість створення словника безпосередньо з тексту або колекції текстів. У цьому розумінні сучасна корпусна лексикографія спирається на традиції **текстоорієнтованих досліджень**. На позначення словників, укладених на матеріалі тексту або корпусу текстів, в українській лінгвістиці традиційно вживається термін текстоорієнтовані [14, с. 14] або текстозорієнтовані [9, с. 75]. На відміну від системоорієнтованих словників, які описують систему мови, текстоорієнтовані відображають закономірності мовлення, функціональні властивості мовних одиниць. При цьому, за В. І. Перебийніс, функціонування розуміється не як виконувана одиницею функція, а як її поведінка в мовленні, тобто сукупність її характеристик в усному чи писемному тексті: частота, сполучуваність, місце в тексті, ступінь реалізації її системних характеристик (наприклад, словозмінних форм), комунікативне призначення, прагматичне чи емотивне навантаження, стилістична забарвленість та ін. [15, с. 138]. Окреслені ознаки мовних одиниць залежать від характеру тексту, функціонального або авторського стилю.

На думку В. І. Перебийніс, до текстоорієнтованих словників відносять: 1) конкорданси; 2) словники мови автора; 3) словники цитат, крилатих виразів; 4) частотні словники, що фіксують наскільки поширеною є одиниця в тексті; 5) словопокажчики, які реєструють позицію та адресу одиниці в тексті [14, с. 52–76]. Відносна легкість опрацювання тексту й укладання сприяла появі перших посеред текстоорієнтованих частотних словників. Перша в українському мовознавстві серія з п'яти частотних словників (художня проза, драма, поезія, наукові тексти, суспільно-політичні тексти) була укладена вручну на вибірці всього 50000 слововживань кожний у 1967 р. [14, с. 58]. Отже, на етапі становлення корпусної лексикографії в українському, як і загалом у радянському мовознавстві 1960-1970 рр. доелектронні, укладені вручну колекції текстів розглядалися виключно як джерело досліджень з лінгвостатистики, зокрема з статистичної лексикографії. Незважаючи на суто корпусне підґрунтя лінгвостатистичних досліджень, слід відзначити пріоритет укладання частотних словників над розробкою електронних корпусів текстів у радянській і, зокрема українській лінгвістиці 1960-1970 рр.

Крім традиції текстоорієнтованих досліджень, до історичних передумов виникнення української корпусної лексикографії слід зарахувати потребу в **автоматизації** трудомістких лексикографічних процесів, таких, як укладання реєстру словника, лексичної картотеки й словникової статті. На думку Л. Засоріної, перші спроби автоматизації лексикографічного аналізу Р. Буза й А. Джіланда мали значний вплив на розвиток приклад-

них досліджень [8, с. 149–150] у радянському мовознавстві в 1960 рр. Однак, відомо, що перший корпусний словник – H. Kučera, W. N. Francis. Computational Analysis of Present-Day American English – було укладено в 1967 р. на матеріалі Браунівського корпусу текстів, використаного Л. Засоріною лише в якості моделі для наступного Частотного словаря русского языка (1977 р.). Подібні тенденції спостерігаються і в українському мовознавстві. Незважаючи на відсутність сформованого корпусного напряму створення машинного фонду й моделювання мовних явищ за допомогою комп'ютера сприяли появі різноманітних частотних словників і конкордансів на основі повнотекстових баз даних [9, с. 95; 6, с. 36–37]. Зокрема традицію укладання текстоорієнтованих словників було продовжено виданням Частотного словника сучасної української художньої прози (1981 р.), створеного на вибірці текстів обсягом у 500 тис. слововживань, як і всі докорпусні словники, вручну [14, с. 58]. Період 1980–1990 рр. характеризується розвитком широкомасштабних лінгвостатистичних досліджень, виконуваних вручну на великих обсягах текстового матеріалу. Безумовно, традиції текстоорієнтованих лінгвостатистичних досліджень 1960–1990 рр. вплинули на методику лексикографічного аналізу, однак в Україні розвиток власне корпусної лексикографії гальмувався через відсутність мотивації до побудови корпусів текстів і недостатню комп'ютеризацію лінгвістичних досліджень. Фактично, докорпусні дослідження, виконувані на доелектронних колекціях текстів, були поштовхом виключно для розвитку української статистичної лексикографії.

Процес формування концептуальних засад сучасної української корпусної лексикографії слід розглядати в аспекті розвитку мовознавчої думки другої половини XX століття. Теоретичною основою текстоорієнтованих досліджень і, зокрема корпусної лінгвістики, безперечно вважається **структуралізм** – система поглядів та методів дослідження, які базуються на розумінні мови як знакової системи з дискретними структурними елементами та використанні формальних прийомів опису [6, с. 14]. У цьому розумінні, поява першого електронного корпусу текстів (Brown Corpus, 1963 р.) мотивується домінуванням в північноамериканській традиції доктрини дескриптивної лінгвістики, «в якій більше, ніж у інших напрямах структуралізму, виявляється тенденція до використання ймовірнісних та статистичних методів дослідження» [6, с. 16]. Власне терміносполука «корпусна лінгвістика» з'явилась значно пізніше у 1980 рр. : корпусні лінгвісти того часу називали себе структуралістами [12, с. 273] й наслідували принципово структурний підхід до вивчення мови [6, с. 14].

З іншого боку, на думку В. Тойберта, історично поява корпусної лінгвістики була відповіддю на необхідність вирішення лінгводидактичних проблем англійської мови як іноземної [20, p. 137–138]. З огляду на вище сказане можна стверджувати, що корпусна лінгвістика мала чітко визначене місце в британському контексті **прикладної лінгвістики** з акцентом на викладання мови та укладання словників. Саме під впливом теоретико-методологічних засад британського лексикографа Е. С. Хорнбі (Oxford Advanced Learner's Dictionary, 1947 р.) в середині 1990 – на початку 2000 рр. в українському словникарстві з'являються праці **з навчальної лексикографії**, які суміщають риси корпусних і системоорієнтованих словників: Англо-український та українсько-англійський словники (1995-2005 рр.) [16, с. 106–107] і Słownik rosyjsko-ukraińsko-polski (2013 р.) [19]. Однак наведені вище навчальні словники можна лише умовно вважати корпусно-базованими, оскільки при їх укладанні корпусний підхід поєднано з інтуїтивним: зокрема, включення слів до загального реєстру словника здійснено на базі частотних характеристик, отриманих у результаті дослідження текстового ресурсу, а добір їхніх вживань [16, с. 106–107] або перекладів [19] базується на інтуїції укладачів.

Початок етапу **власне корпусної лексикографії**, базованої на електронних корпусах текстів, припадає в українському мовознавстві на перше десятиріччя XXI ст. При цьому перевага віддається укладанню корпусних частотних словників, а не побудові корпусів текстів. Подібні тенденції розвитку корпусної лексикографії є характерними й для інших країн пострадянського простору. Однак російські лексикографи мали змогу користуватись укладеним шведськими русистами електронним Упсальським корпусом [13, с. 197]. Українська мова залишалась однією з небагатьох, що не мали репрезентованого у вільному доступі національного корпусу, його створення лише усвідомлювалось як нагальне завдання й перспектива розвитку української корпусної лексикографії. Отже, з одного боку, занепад кібернетики в СРСР завадив корпусній революції в українській лексикографії, а з іншого – зазначені особливості розвитку корпусної лінгвістики сприяли усвідомленню зарубіжного досвіду побудови сучасних електронних корпусів. На відміну від світової традиції в українській лінгвістиці поява праць, що обґрунтовують принципи побудови та застосування електронних корпусів [6; 9, с. 74–103] і корпусних словників [4] збігається за хронологією. Здійснене дослідження дозволяє виділити три етапи розвитку української корпусної лексикографії і встановити відповідні типи корпусних словників: 1. Статистична лексикографія (1960–1990 рр.) – докорпусні частотні словники, конкорданси. 2. Докорпусна навчальна лексикографія (1990–2010 рр.) – докорпусні навчальні словники. 3. Власне корпусна лексикографія (з 2004 р.) – корпусні словники.

На сьогодні, українська корпусна лінгвістика представлена у вільному для користувача доступі двома дослідницькими корпусами текстів української мови [22; 23], Навчальним корпусом англійських текстів – Ukrainian Corpus of Learner English (UCLE) [24] і Багатомовним паралельним корпусом усного мовлення [25]. Посеред представлених у вільному доступі текстових ресурсів новітній Корпус текстів української мови укладено колективом кафедри української мови та прикладної лінгвістики Донецького національного університету з метою вивчення проблеми граматичної службовості [3, с. 224]. У межах проекту реалізовано технічні й програмні аспекти реалізації корпусу, розроблено морфорозмітку й метарозмітку, а також систему тегів для службових частин мови. На сьогодні Корпус текстів української мови загальним обсягом близько 5 млн. слововживань функціонує в тестовому режимі [3, с. 224–225]. Найбільший за обсягом дослідницький Корпус сучасної української мови [23] побудовано як інформаційно-довідкову систему, призначену для з'ясування різних питань вивчення української мови. Корпус загальним обсягом у 13 млн. словоформ анотовано за якіс-

ними й кількісними ознаками різних мовних одиниць на рівні морфеміки, морфології й синтаксису, а також забезпечено пакетами програм для укладання електронних картотек і параметризованої бази даних, на базі корпусу розроблено серію корпусних словників [5, с. 46–47].

Навчальні корпуси представлені в українській корпусній лінгвістиці тестовою версією корпусу англійських текстів UCLE [24], створеною в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету. Загальний обсяг текстів студентських есе становить понад 180 тис. слововживань [12, с. 30]. Програмне забезпечення навчального корпусу дозволяє будувати повні конкордансні списки та за ключовим словом, здійснювати пошук окремих слів і словосполучень, сортувати списки слів, відображати знайдені словоформи у необмеженому контексті, отримувати статистичну інформацію про окремі елементи корпусу. Багатомовні корпуси текстів в українській корпусній лінгвістиці представлені паралельним корпусом усного мовлення [25]. Корпус загальним обсягом біля 8 млн. розроблено в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету на базі субтитрів серіалів комедійного, драматичного й науково-популярного жанру. Аналізований корпус включає підкорпуси оригінальних текстів англійською мовою загальним обсягом біля 2 млн. та відповідних перекладів німецькою – 0,65 млн., французькою – 0,8 млн., українською – 0,2 млн., російською – 1,1 млн., іспанською – 1,2 млн. і грецькою – 1,2 млн. Особливістю розробки даного паралельного корпусу текстів є вирішення проблеми автоматичного вирівнювання речень через використання параметру синхронізації часу появи субтитрів на екрані. Програмне забезпечення корпусу дозволяє здійснювати пошук перекладних еквівалентів слів і словосполучень у контексті речення, однак морфологічне анотування й модуль лематизації відсутні.

Доступність зазначених корпусів текстів і гнучкість програмного забезпечення дозволяють прогнозувати швидкий розвиток корпусної методології як «підґрунтя повного опису мовних явищ, нездійсненого в докорпусний період» [18, p. 117], у тому числі й лексикографічного аналізу. В даному дослідженні процес укладання корпусного словника розуміється як здійснення корпусного дослідження. Саме тому, в основу встановлення типології сучасних корпусних словників покладено триступеневу систематику корпусних досліджень на базі методології [18, p. 115]. За зазначеною систематикою розрізняють: а) корпусно-інформативні дослідження з використанням корпусу як колекції природномовних ілюстрацій на підтвердження заздалегідь сформульованих гіпотез дослідника; б) корпусно-базовані – з повним аналізом корпусу за кількісними та якісними параметрами на основі апріорно сформульованих теоретичних припущень; в) корпусно-керовані дослідження з повним генеруванням моделі або побудови теорії мови з корпусу текстів. Використання окресленої систематики для встановлення типології корпусних словників свідчить про те, що на сьогодні українська корпусна лексикографія представлена всіма типами словників.

– **Корпусно-інформативні словники:** Тримовний тлумачний словник термінів з комп'ютерної лінгвістики з ілюстрацією вживання в англо-українсько-російському корпусі текстів [16; 25].

– **Корпусно-базовані словники:** Текстозорієнтований тезаурус лінгвістичних термінів з верифікацією на корпусі текстів з різних розділів лінгвістики [17, p. 70–71], Морфемні й словотвірні словники Корпусу української мови [5, с. 47], Комп'ютерний фонд інновацій [10, с. 26], Словник часток [7, с. 21].

– **Корпусно-керовані словники:** 1) Частотні – Алфавітно-частотні словники, словник-конкорданс Корпусу української мови [26], Частотні словники паралельних текстів [1, с. 158]; 2) словники мови авторів – Словники поетів [5, с. 47], Частотный словарь избранной поэзии И. Бродского [2, с. 9]; 3) словники неолексем, синонімів, антонімів, фразеологізмів Корпусу української мови [5, с. 47]; 4) словники синтаксичних моделей керування [5, с. 47], актуальний англомовний словник українських есе [12, с. 30].

Завдяки потужній лінгвостатистичній традиції в українському мовознавстві, значний доробок корпусно-керованих досліджень становлять частотні словники [1; 2; 26]. Здійснене дослідження етапів становлення й основних тенденцій розвитку української корпусної лексикографії дозволяє дійти таких висновків:

1. У сучасному мовознавстві корпусна лексикографія визначається як лінгвістична дисципліна, яка вивчає теорію і практику укладання корпусних словників.

2. Корпусна лексикографія базується на понятті корпусу текстів, під яким розуміється електронний ресурс, використовуваний в якості лексикографічного або лексичного джерела для укладання певного словника.

3. Передумови виникнення української корпусної лексикографії становлять традиції лінгвостатистичних досліджень й автоматизація лексикографічного аналізу.

4. Концептуальною основою української корпусної лексикографії є структуралізм, зокрема, положення американської дескриптивної лінгвістики й теоретико-методологічні засади британської прикладної лінгвістики.

5. Розвиток української корпусної лексикографії включає три етапи: статистичної лексикографії – докорпусних частотних словників і конкордансів, докорпусної навчальної лексикографії – докорпусних навчальних словників і власне корпусної лексикографії – корпусних словників.

6. Основними тенденціями розвитку української корпусної лексикографії є потужна традиція лінгвостатистичних досліджень, пріоритет укладання частотних словників над електронними корпусами текстів і традиція корпусно-базованої навчальної лексикографії.

7. Встановлення типології корпусних словників на базі методології дослідження доводить: сучасна українська корпусна лексикографія представлена корпусно-інформативними, корпусно-базованими й корпусно-керованими словниками.

Література:

1. Бобкова Т., Перебийніс В., Сорокін В. Частотні словники паралельних текстів / Т. Бобкова, В. Перебийніс, В. Сорокін // Людина. Комп'ютер. Комунікація : [зб. наук. праць]. – Львів : Вид. Національного університету «Львівська політехніка», 2008. – С. 158–160.

2. Бобкова Т. Составление частотного словаря избранной поэзии Иосифа Бродского / Т. Бобкова // Комп'ютерна лінгвістика : сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 9–13.

3. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії : [зб. наук. праць]. – Вип. 26. – Донецьк : ДонНУ, 2013. – С. 224–229.

4. Дарчук Н. П. Структурно-статистическая база данных современного украинского языка на основе частотных словарей / Н. П. Дарчук // Слово и словарь = Vocabulum et vocabularium : [сб. науч. тр. по лексикографии]. – Гродно : ГрГУ, 2005. – С. 194–196.

5. Дарчук Н. П. Дослідницький корпус української мови : основні засади і перспективи / Н. П. Дарчук // Вісник Київського національного університету ім. Тараса Шевченка. Серія : Літературознавство. Мовознавство. Фольклористика. – К. : ВПЦ «Київський університет», 2010. – № 21. – С. 45–49.

6. Демська-Кульчицька О. Основи національного корпусу української мови : [монографія] / Орися Демська-Кульчицька. – К. : Інститут української мови НАНУ, 2005. – 219 с.

7. Загнітко А., Ситар Г., Данилюк І. Структура і модель бази даних «українські частки та їхні еквіваленти»/ А. Загнітко, Г. Ситар, І. Данилюк // Комп'ютерна лінгвістика : сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 21–22.

8. Засорина Л. Н. Письмо в редакцию / Л. Н. Засорина // Вопросы языкознания. – М. : Изд. «Наука», 1968. – № 6. – С. 149–150.

9. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики : [підручник] / Євгенія Анатоліївна Карпіловська. – Донецьк : Юго-Восток, ЛТД, 2003. – 188 с.

10. Карпіловська Є. Комп'ютерне моделювання мовних змін : система мови і текст / Є. Карпіловська // Комп'ютерна лінгвістика : сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 25–26.

11. Коломієць В., Котик С. Спеціальний навчальний корпус текстів UCLE : сучасний стан і перспективи використання / В. Коломієць, С. Котик // Комп'ютерна лінгвістика : сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 29–32.

12. Лендау С. І. Словники : мистецтво та ремесло лексикографії / Сидні І. Лендау; [пер. з англ.]. – К. : К. І. С., 2012. – 480 с.

13. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. Национальный корпус русского языка как инструмент лексикографа / О. Н. Ляшевская, В. А. Плунгян, Д. В. Сичинава // Слово и словарь = Vocabulum et vocabularium : [сб. науч. тр. по лексикографии]. – Гродно : ГрГУ, 2005. – С. 197–202.

14. Перебийніс В. І., Сорокін В. М. Традиційна та комп'ютерна лексикографія : [навч. посібник] / Валентина Ісидорівна Перебийніс, Віктор Михайлович Сорокін. – К. : Вид. Київського національного лінгвістичного університету, 2009. – 218 с.

15. Перебийніс В. І. Системні та функціональні характеристики мовних одиниць / В. І. Перебийніс // Вісник Харківського національного університету ім. В. Н. Каразіна. – Харків : Константа, 2004. – № 635. – С. 138–141.

16. Bobkova T. etc. Corpus of computational linguistic texts / T. Bobkova etc. // Computer Treatment of Slavic and East European Languages. – Bratislava : Tribun, 2009. – P. 35–40.

17. Darchuk N. P., Sorokin V. M. Text-Oriented Thesaurus Retrieval Sysytem for Linguistics / N. Darchuk, V. Sorokin // Computer Treatment of Slavic and East European Languages. – Bratislava : Tribun, 2009. – P. 65–77.

18. Mukherjee J. The state of the art in corpus linguistics : three book-length perspectives / J. Mukherjee // English Language and Linguistics. – Vol. 8. 1. – Cambridge : Cambridge University Pres, 2003. – P. 103–119.

19. Świkszcz-Kobyłecka M., Bobkowa T. Słownik rosyjsko-ukraińsko-polski / Marioła Świkszcz-Kobyłecka, Tatiana Bobkowa. – Toruń-Kijów : MARTOM, 2013. – 73 s.

20. Teubert W. Linguistique de corpus : un alternative / W. Teubert // Semen. Critical Discourse Analysis I. Les notions de contexte et d'acteurs sociaux / par A. Petitclerc, Ph. Schepens. – Vol. 27. – Presses Universitaires de Franche Comté, 2009. – P. 130–152.

Джерела ілюстративного матеріалу:

21. Багатомовний паралельний корпус усного мовлення. – К. : КНЛУ, 2010. – Режим доступу : http://www.complinguide.com.ua/Corpus. aspx

22. Корпус текстів української мови кафедри української мови та прикладної лінгвістики Донецького національного університету. – Режим доступу : http://corpora.pp.ua/bonito/

23. Корпус текстів української мови. – Режим доступу : http://www.mova.info/corpus.aspx?l1=209

24. Навчальний корпус текстів UCLE. – Режим доступу : http://www.complinguide.com.ua/Ucle_index.aspx

25. Тримовний тлумачний словник термінів з комп'ютерної лінгвістики – Режим доступу : http://www.complinguide.com.ua/Glossary.aspx

26. Частотні словники. – К. – Режим доступу : http://www.mova.info/Page.aspx?l1=57