

Є.В. Купріянов

Комп'ютерна лексикографія як проблема сучасного мовознавства
(історичний аспект)

Одним із важливих напрямків сучасного вітчизняного мовознавства є комп'ютерна лексикографія. Сьогодні перед дослідниками постають важливі питання, що стосуються теоретичних і практичних аспектів укладання комп'ютерних словників, наукова значущість яких є безсумнівною. Необхідним етапом у вирішенні цих питань є осмислення особливостей формування цього розділу мовознавчої науки – її передумов, методологічної бази, напрямків наукового пошуку.

Стаття присвячена висвітленню деяких аспектів історичного розвитку вітчизняної та зарубіжної лексикографії.

Завдання статті – розглянути основні етапи розвитку комп'ютерних технологій укладання словників та визначити чинники, що зумовили появу такого напрямку у мовознавстві, як комп'ютерна лексикографія.

Поява комп'ютерів активно вплинуло на розвиток лексикографії. Спочатку вони використовувалися для підготовки паперових словників, інакше кажучи слугували печатною машинкою. Але згодом виявилось, що комп'ютери можуть виконувати такі функції, як редагування, зберігання будь-якої лексикографічної інформації, і тому з'явилися комп'ютерні корпуси текстів, а далі й машинозчитувані словники.

Однією за перших розробок у зазначеній галузі став комп'ютерний корпус сучасного американського варіанту англійської мови, або просто Брауновський корпус (*Brown University Standard Corpus of Present-Day American English*, або просто *Brown Corpus*), який було укладено у 1960 році в університеті Брауна дослідниками Генрі Кучера та Нельсоном Френсісом.

Брауновський корпус являв собою ретельно укладений матеріал з американської англійської мови обсягом 1 млн. слів, підібраних з багатьох джерел. Кучера і Френсіс зробили на цьому матеріалі різноманітні комп'ютерні аналізи, завдяки яким науковці отримали багатий та різноманітний твір, який увібрав у себе елементи лінгвістики, психології,

статистики та соціології. Корпус широко використовувався у комп'ютерній лінгвістиці і протягом багатьох років був одним із самих популярних ресурсів у цій галузі [4].

Згодом було зроблено декілька спроб створити корпуси більшого розміру. У Великобританії такими проектами були Банк Англійської мови (Bank of English) и Британський Національний Корпус (British National Corpus, BNC).

Разом з цим, комп'ютерна лексикографія також розвивалася в галузі обробки природномовних текстів. Дослідженнями у цій галузі вела лабораторія автоматичної обробки документів і лінгвістики, яка була створена у 1966 році (Laboratoire d'Automatique Documentaire et Linguistique – LADL) при Університеті Марни (Франція) і переросла в центр європейської мережі RELEX. Дослідницька програма LADL спрямована на розробку фундаментальних інструментів з обробки природномовних текстів. Її інструментарій складають:

- лінгвістичні компоненти: електронні словники і граматики, в основному для англійської, французької, іспанської і корейської мов;
- програмне забезпечення алгоритмів, які працюють зі словниками і грамами на корпусних масивах текстів для того, щоб визначити і лемматизувати (звести до глибинних форм) повнозначні уривки текстів; при цьому основним додатком є автоматичне індексування текстів, інформаційний пошук у повних текстах, допомога при перекладі [1, с. 17].

До електронних словників, розроблених LADL, належать багатомовні словники флексій DELAF. Такий словник містить близько 600 тис. словоформ для простих слів і близько 150 тис. словоформ для складних слів (DALACF). Основні слова в цих словниках наведені як автомати кінцевих станів. Це дає можливість здійснити потужне індексування текстів. Просте слово подано як автомат кінцевого стану, що містить морфологічну і синтаксичну інформацію [1, с. 18].

Окремим напрямком комп'ютерної лексикографії стало створення машинних словників для інформаційно-пошукових систем (таких як GAT, що використовувалася з 1966 року комісією з атомної енергетики США й у Євроатомі для отримання інформації з робіт російських атомщиків) і систем машинного перекладу (наприклад, система SYSTRAN, що була створена у 70-х роках для корпорації General Motors для прискорення процесів перекладу матеріалів).

Перші комп'ютерні словники, що являли собою машинні версії паперових словників, з'явилися у кінці 70-х та початку 80-х років минулого століття. Вони слугували дослідникам як зручний матеріал для лексикографічних досліджень.

У Великобританії було створено такі машинні версії традиційних словників англійської мови, як *Oxford Advanced Learner's Dictionary (OALD)*, *Dictionary of Contemporary English (LDOCE)* і *Collins Cobuild English Dictionary (COBUILD)*.

Словник *Oxford Advanced Learner's Dictionary (OALD)* став доступним у машинозчитуваній формі наприкінці 1970-х. Комп'ютер не грав ніякої ролі під час лексикографічної підготовки словника. В основному, це була комп'ютерна перфокарта. Це був перший машинозчитуваний словник на перфокарті [5, с. 25].

На початку 1980-х років з'явився машинний словник *Longman Dictionary of Contemporary English (LDOCE)*. Автори під час його підготовки використовували комп'ютерні засоби перевірки послідовності дефініцій слів. Словник *LDOCE* став першим машинозчитуваним словником, створеним з використанням комп'ютера [5, с. 25–26].

Словник *COBUILD* являв собою перший машинозчитуваний словник, розроблений за допомогою комп'ютера. Розробка словника складалася з чотирьох етапів: збір даних, відбір словникових статей, зіставлення дефініцій для словникових статей і упорядкування словникових статей. Комп'ютер

також використовувався для перевірки послідовності та повноти словникових статей [5, с. 26].

Першим словником, створеним на базі ЕОМ, став 15-томний «Словник французької мови» (8 тис. словникових статей). Але це була лише частина автоматичної картотеки слів із прикладами, що охоплювала корпус текстів у 90 млн. слововживань. Іншими словами, вже на початковому етапі розробок електронного словника лексикографи ставили задачі більш широкі, ніж рутинне перекладання словника на магнітні носії. Майже одночасно з автоматизацією цих задач комп'ютерна лексикографія перейшла до вирішення проблем іншої якості. З'явилася реальна перспектива використання результатів лексичного аналізу слів, що багато в чому перевершував можливості «ручного» лексикографування, тобто роботи окремих дослідників з аналізу великих мовних масивів для створення словників. стандартного програмного забезпечення для персональних комп'ютерів мільйонів користувачів [1, с. 9].

У свою чергу, із виникненням машинозчитуваних словників стало можливим їх видання на CD-ROM, тобто на лазерних дисках. Текст першого видання Оксфордського словника став доступним у 1988 році. Згодом, з'явилися три електронних версії другого видання. Перша версія (1992 р.) була за змістом ідентичною до паперового аналога, але сам диск був незахищеним від копіювання. Друга версія (1999 р.) мала деякі доповнення й удосконалене програмне забезпечення та більш зручні засоби пошуку, але існували недоліки у захисті від копіювання. Третя версія (2002 р.) містить додаткову кількість слів та більш досконале програмне забезпечення, хоча й досі залишалися недоліки у захисті копіювання, як і в попередніх версіях. Інтернет-версія Оксфордського словника стала доступною 14 березня 2000 року [6]

По мірі удосконалення комп'ютерних технологій розширювалися можливості й функції електронних словників. Вони вже слугували не тільки засобами збереження лінгвістичної інформації, складовими елементами

систем обробки мови, машинного перекладу, але й могли виконувати такі функції, як навчання мові, швидкий пошук. Словники набули можливості вибіркового подання інформації, що міститься у словниковій статті, показу декількох словникових статей одночасно, тощо.

Зародження вітчизняної комп'ютерної лексикографії, на нашу думку, припадає на середину 80-х років минулого сторіччя, коли вийшла програма ДКНТ СРСР зі створення національних машинних фондів СРСР.

У рамках вищеназваної програми вченими з кафедри програмного забезпечення та електротехніки «Львівської політехніки» була створена система підтримки багатомовних термінологічних словників під назвою «СЛОВО». У ній відпрацьовані технологічні проблеми підготовки словників до друку. Так, була розроблена комп'ютерна версія англо-українсько-англійського словника термінів інформаційних технологій. Її обсяг – біля 9 тис. слів [1, с. 11].

Розвиток комп'ютерної лексикографії в Україні після набуття незалежності характеризувався процесами інтеграції нашої країни у всевітню інформаційну спільноту та реалізації мовної політики з боку держави, зокрема у комп'ютерно-інформаційній сфері для створення національної словникової бази. Про це свідчать постанова Кабінету Міністрів України від 8 вересня 1997 року «Про затвердження Комплексних заходів щодо всебічного розвитку і функціонування української мови», указ Президента України «Про розвиток національної словникової бази», від 1999 року, розпорядження Кабінету Міністрів «Про першочергові завдання із створення національної словникової бази» від 22 листопада 2000 року, постанова Верховної Ради України «Про функціонування української мови в Україні» від 22 травня 2003 року.

У зв'язку з вищезазначеними процесами, все більше виникає необхідність у створенні автоматизованих лексикографічних систем – комп'ютерних словників, тезаурусів, програм обробки природномовних текстів, які б могли забезпечувати автоматизоване й машинне редагування,

інформаційний пошук, розпізнавання і корегування текстів, укладання нових словників, накопичення і підтримки матеріалів для електронних бібліотек тощо.

Словосполучення «національна словникова база» вперше офіційно з'явилося в тексті Указу Президента України від 07.08.1999 р. №967 «Про розвиток національної словникової бази». Хоча у зазначеному документі не було сформульовано визначення цього поняття, проте певні ключові слова, асоційовані з ним, наводилися. А саме: термін *національна словникова база* пов'язувалася із розширенням сфери функціонування української мови, створенням нового покоління академічних україномовних словників та їх електронних відповідників для комп'ютерних інформаційних систем (проект «Словники України»). Реалізація цього проекту покладалася на Національну академію наук України [3, с. 301].

Першим, хто почав проводити активні наукові дослідження та розробки у галузі формування національної словникової бази для української мови, на нашу думку, став Український мовно-інформаційний фонд, який було засновано у 1991 році.

Тут було створено базові моделі комп'ютерної лексикографії, які склали підґрунтя для розробки відповідних технологій, започатковано серію лексикографічних праць нового покоління – «Словники України» [3, с. 27].

Спочатку ця організація займалася питаннями лексикографії, а також складанням спеціалізованих словників, що виходили в серії «Словники України».

При підготовці видань велика увага приділялася автоматизації процесу складання, для чого був проведений ряд фундаментальних досліджень з комп'ютерної лінгвістики. Так, ще для першого орфографічного словника розробили формальну теорію класифікації способів словозміни в українській мові та відповідну алгоритмічну базу, що дозволила на основі вихідної форми автоматично одержувати всі словоформи. Спочатку були визначені 280 парадигматичних класів, але сьогодні більш глибокі дослідження дали

можливість вивести їх близько 1,5 тис. і в такий спосіб охопити весь словниковий запас літературної української мови [Черницький, с. 12].

Серед завдань Мовно-інформаційного фонду своєю масштабністю виділяється проект зі створення фундаментальної багатотомної академічної лексикографічної системи «Словник української мови». У 2001 році випущено перший повномасштабний український словник на лазерному диску у вигляді інтегрованої лексикографічної системи "Словники України", яка містить унікальний набір словникових функцій: на реєстрі 152 тис. одиниць тут унаочнено повну словозмінну парадигму і транскрипцію згідно з правилами української орфографії та орфоепії; в системі подано понад 56 тис. фразеологізмів, близько 9 тис. синонімічних рядів, понад 2,1 тис. антонімічних пар. За своїми лінгвістично-інформаційними параметрами система «Словники України» не має аналогів у світі і є незамінною в комп'ютерному діловодстві, навчанні української мови, редакційно-видавничій діяльності та проведенні лінгвістичних досліджень. З метою популяризації досягнень вітчизняної лінгвістики було проведено низку презентацій випусків серії «Словники України», які одержали значну пресу й широко висвітлювалися засобами масової інформації. Понад 20 тис. примірників видань цієї серії було безкоштовно передано закладам освіти, науки, бібліотекам, міністерствам та відомствам України.

Упродовж 1998-2003 рр. фондом було розроблено системотехнічні засади створення та ведення лінгвістичних корпусів, на основі чого сформовано комп'ютерні лексикографічні й текстові масиви української мови загальнонаціонального значення, серед яких слід відзначити:

- лінгвістичний корпус українських текстів, призначений для постійного формування, збереження та використання художньої літератури (від І.Котляревського до XXI ст.), наукової та науково-популярної, суспільно-політичної, публіцистичної літератури (у тому числі перекладної);
- фундаментальну електронну лексичну картотеку обсягом понад 30 млн. слововживань;

- лексикографічні бази даних понад 20-ти словників (тлумачного, орфографічного, орфоепічного, фразеологічного, синонімічного, антонімічного, граматичного та ін.);
- систему природномовного індексування українських текстів та баз даних;
- автоматизовану систему конверсії текстів словників до комп'ютерних лексикографічних баз даних;
- мережевий інструментальний комплекс для підтримки сучасної цифрової технології створення фундаментальних лексикографічних праць [2].

Не менш важливу роль у розвитку комп'ютерної лексикографії зіграла кафедра Комп'ютерної лексикографії Інституту української мови імені О.О. Потебні, яку було засновано згідно з постановою Кабінету Міністрів України від 8 вересня 1997 року. До широкомасштабних проєктів, започаткованих кафедрою, відносяться:

- національний корпус української мови з планованим початковим обсягом у 2 мільйони 500 тисяч слів, що являє собою систематизоване, структуроване, програмно оброблене зібрання взірцевих текстів української мови всіх варіантів та форм її існування. Призначений для лінгвістичних досліджень та технологічних застосувань;
- електронна лексична картотека, сукупність лексичних карток із заголовними словами, текстами ілюстраціями вживання цих слів у відповідному значенні та вказівкою на джерело ілюстративного тексту. Це аналогічна до традиційної сукупність лексичних карток, але зберігається вона на електронних носіях, містить додаткові інформаційні поля, і з неї за алгоритмом лематизації автоматично формується заданий реєстр.

Лабораторія комп'ютерної лінгвістики Інституту філології Київського національного університету розробила такі лексикографічні продукти під керівництвом Дарчук Н.П.:

- автокоректор РУТА, за допомогою якого можна перевіряти правопис, тобто автоматично знайти і виправити помилки у словах, виконати граматичний, пунктуаційний та стилістичний контроль, розставити переноси під час форматування тексту, скористуватися словником синонімів;

- автоматизована система українсько-російського перекладу ПЛАЙ, що володіє лексико-граматичним аналізом мов і дає в розпорядження користувача могутні можливості для обробки текстів будь-якої складності.

Для даного етапу розвитку комп'ютерної лексикографії, як влучно помітив В.А. Широков, характерна «...розробка стаціонарних комп'ютерних зібрань лексикографічної інформації у відповідних мовах та сукупностях мов. Інформаційною основою всіх словникових систем...стали ретельно опрацьовані філологами формати запису, які містили певним чином систематизовану й класифіковану граматичну, морфологічну, фонетичну, семантичну, фразеологічну, етимологічну та іншу лінгвістичну інформацію про лексичні одиниці. Такі словникові системи створили необхідне підґрунтя й відіграли відповідну практичну роль у створенні й випуску нових поколінь національних словників, у формуванні стандартних лексикографічних масивів національних мов та проведених на цій основі комп'ютерних лексикографічних та лексикологічних дослідженнях» [Цит. за 1, с. 20].

На сучасному етапі розробками у сфері комп'ютерної лексикографії, окрім зазначених вище установ, стали також займатися комерційні організації. До таких організацій можна віднести *АВВУУ Україна* (м. Київ), *Софтпром* (м. Київ), *Лінгвістика 93* (м. Харків), *Медіком*, *Proling Office*. В основному вони спеціалізуються на укладанні перекладних словників, що містять загальну і спеціальну лексику, автоматичні системи перекладу, такі як, наприклад, *АВВУУ Lingvo 12*, *РУМП*, *ПАРС*, *GEISHA*, *PRAGMA* та ін.

Таким чином, на початкових етапах розвитку комп'ютери у галузі лексикографії використовувалися як друкарська машинка, яка могла замінити, витирати літери, слова й навіть цілі частини текстів. А з виникненням операційних систем, таких як *Windows*, комп'ютер надав

лексикографу не тільки інструментальні засоби редагування, але й форматування, створення оригінал-макету словника.

Поява планшетних сканерів та програм розпізнавання значно прискорило процес набору та редагування тексту. Водночас було створено програмне забезпечення, що дозволяло здійснювати пошук, індексацію текстів. Це привело до виникнення лінгвістичних баз даних, електронних бібліотек й картотек.

Автоматизовані лексикографічні бази у вигляді електронних словників зараз становлять невід'ємну частин систем машинного перекладу, інформаційного пошуку, редагування та правки текстів, а також обробки великих текстових масивів та їх зберігання як окремої задачі створення електронних бібліотек.

Комп'ютерні словники на оптичних носіях дали змогу перекладачам, науковцям швидко знаходити будь-яку інформацію про слово таку, як переклад, орфографічну, граматичну інформацію, тлумачення тощо.

АННОТАЦІЯ

В статье рассматривается история развития зарубежной и отечественной компьютерной лексикографии.

SUMMARY

The present article is devoted to historical development of computer lexicography abroad and in Ukraine.

ЛІТЕРАТУРА

1. Комп'ютерна лексикографія: [Навч. посібник] / Черницький В.Б. – Нац. ун-т кораблебудування ім. адм. Макарова. – Миколаїв: НУК, 2004.
2. Український лінгвістичний портал // <http://www.ulif.com.ua/about/aboutpage.php>.
3. Широков В.А. Феноменологія лексикографічних систем. – К.: Наукова думка, 2004.
4. Brown Corpus // http://en.wikipedia.org/wiki/Brown_Corpus.
5. C. Magee Computational Lexicography B.A. (Mod.) CSLL Final Year Project, May 2000, Supervisor: Dr. Carl Vogel.
6. Oxford English Dictionary http://en.wikipedia.org/wiki/Oxford_English_Dictionary#Electronic_versions.