В. Балог, к.філол.н.[*]
Інститут української мови НАН України (Київ)
УДК 81’33

ПЕРЕДУМОВИ СТВОРЕННЯ КОРПУСУ СЛОВНИКІВ

У статті досліджено лексикографічний аспект у створенні корпусу мови. Мета проекту полягає у створенні локальної джерельної бази для будь-яких досліджень на основі словників. Вважаємо, що корпусний підхід до вирішення поставленого завдання дозволить не тільки уніфіковано формалізувати дані словників, але й ввести в систему будь-який словник, незалежно від ступеня його готовності.

Сьогодення ставить проблему економії часу. Ця теза не нова, проте вимагає створення умов для ефективності робочого процесу в будь-якій сфері суспільної діяльності. Ця ж теза провокує розвиток корпусної лінгвістики як шлях до створення „робочого місця” лінгвіста, полегшення процесу накопичення фактичного матеріалу та його систематизації. Необхідність створення корпусу текстів природної мови не потребує додаткових доказів, оскільки не тільки значно полегшує дослідницький процес, а й у суто лінгвістичному аспекті є репрезентантом мови: її синхронії та діахронії. На наше глибоке переконання корпус має бути публічним об’єктом, загальнодоступним для користувача.

Окремим аспектом функціонування корпусу вважаємо лексикографічний. Створення корпусу словників – це науково-дослідний проект, втілення якого зумовлене потребою зібрання всього масиву лексикографічних джерел української мови в електронному вигляді в одній площині з однаковим доступом до зафіксованої в них інформації. Мета проекту полягає у створенні локальної джерельної бази для будь-яких досліджень на основі словників. Корпусний підхід до вирішення поставленого завдання дозволить не тільки уніфіковано формалізувати дані словників, але й ввести в систему будь-який словник, незалежно від ступеня його готовності (це дуже актуально, оскільки деякі українські словники мають незавершений вигляд через історичні обставини), а також у співпраці з програмістами забезпечити зручну пошукову систему за всіма параметрами, представленими у словнику. Особливим завдання ставимо введення в корпус усіх видань окремо взятої праці, що дозволить накопичити фактичний матеріал для досліджень діахронії в межах одного словника.

Сучасна лексикографія поступово вступає в якісно новий етап свого розвитку і в цьому процесі перебуває в перехідному періоді, коли паралельно співіснують так звані традиційні (паперові) словники, їх комп’ютерні варіанти, та власне комп’ютерні словники. Також у цей період вчені застосовують різні – від традиційних до сучасних комп’ютерних – методики укладання словників. Безперечно, майбутнє словникарства за новітньою методикою з огляду не лише на зручність процесу, а передовсім на досконалий результат, динамічність словника, а отже і постійну адекватність стану мови.

Паралельно досить поширеним методом є використання текстів існуючих словників для формування баз даних, які в свою чергу є основою для формування електронних варіантів словників[1]. „Комп’ютерний словник і комп’ютерний варіант традиційного словника становлять результати двох напрямків роботи в сучасній комп’ютерній лексикографії...” [3: 53]. Ми вважаємо таку практику якраз свідченням перехідних процесів, що неодмінно виведуть лексикографію на новий, корпусний рівень. Це необхідний етап розвитку словникарства, на якому закріплюються в електронному вигляді надбання лексикографічної теорії та практики, а також закладаються і вибудовуються основи понятєвого та процедурного апарату для створення справді електронних словників. У руслі останнього методу в комп’ютерній лексикографії, на нашу думку, повинні мати місце дві загальні тенденції на шляху до створення електронного масиву словників: синхронна та діахронна. Синхронний аспект комп’ютерного оброблення лексикографічних джерел полягає у опрацюванні із застосуванням комп’ютерних технологій сучасних словників різних типів. Вони функціонують або окремо у вигляді електронної „книжкової полиці” – кожен словник має

---

[*] *©В.Балог, 2006*

[1] Докладніше про це див: Карпіловська Є.А. Вступ до комп’ютерної лінгвістики. – Донецьк: ТОВ „Юго-Восток, Лтд”, 2003. – С.53–72.

окрему оболонку[1], або у вигляді інтегрованої лексикографічної системи[2], яка має зведений словник з інформацією про різні характеристики слова (морфологічні, правописні, лексико-семантичні, тощо), або як база даних для створення власне електронних лексикографічних об'єктів. Діахронний аспект передбачає створення масиву словників, що вже стали історичним надбанням національної лінгвістичної традиції.

В українській науці розвиток має поки що лише синхронний напрям комп'ютерної лексикографії. Основну увагу приділяють створенню лексичних або лексикографічних масивів з використанням словників ІІ половини ХХ століття (насамперед це академічні словники, створені на основі лексичної картотеки Інституту української мови НАН України). На таку роботу націлені, зокрема, співробітники відділі структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні, де створено Морфемно-словотвірний фонд української мови. Генеральний реєстр слів української мови зведений за матеріалами 5 словників: *Словника української мови в 11-ти томах* (К., 1970-1980 рр.), словника-довідника І. Т. Яценка у 2-х томах „*Морфемний аналіз*" (К., 1980-1981), „*Частотного словника сучасної української художньої прози*" у 2-х томах (К., 1981), „*Словника іншомовних слів*" за ред. О. С. Мельничука (К., 1974) та орфографічної частини „*Словника-довідника з правопису та слововживання*" С. І. Головащука (К., 1989). Сучасні українські словники також лягли в основу досліджень з прикладної (комп'ютерної) лінгвістики науковців і студентів Інституту філології КНУ ім. Тараса Шевченка, Київського національного лінгвістичного університету, Національного університету „Львівська політехніка", Донецького національного університету та інших навчальних закладів України, результатом яких є функціональні навчальні словники різних типів і посібники, відкриті словники, тезауруси, представлені в мережі Інтернет[3].

Важливим процесом у цьому напрямі сучасної лексикографії є теоретичне обґрунтування та практичне втілення теорії лексикографічних систем, який здійснює колектив співробітників Українського мовно-інформаційного фонду НАН України. Базовим здобутком цієї роботи є видана в електронному вигляді Інтегрована лексикографічна система „Словники України", яка також функціонує в режимі on-line на сайтах Український лінгвістичний портал (http//ulif.org.ua) та Нова мова (www.novamova.com.ua), в основу якої покладено „*Орфографічний словник української мови*" (К., 2002), „*Орфоепічний словник української мови*" (К., 2001-2003); два видання „*Фразеологічного словника української мови*" (К., 1993; 2-ге вид. – К., 1999); „*Словник синонімів української мови*" (К., 1999-2000); „*Словник антонімів української мови*" Л. М. Полюги (К., 1999); у перспективі – залучення *Словника української мови у 11-ти томах* (1970-1980 рр.). Звичайно, вказана лексикографічна система не зовсім відповідає всім запитам користувача (передовсім, у частині семантичної характеристики слова), проте зорієнтована на розбудову та вдосконалення.

У вигляді електронної „книжкової полиці" українські словники можна знайти на сайтах www.novamova.com.ua, w. (обіжний аналіз використання інтернет-ресурсу для функціонування українських словників показав, що робота упорядників лексикографічного матеріалу на сайтах, де представлені лексикографічні джерела української мови спрямована на сучасний період українського словникарства[4]. На пошуковому сайті domivka.net у розділі „українська мова" подано наступні адреси: **Нова мова** Освітньо-інформаційний Інтернет-проект оновлення української мови http://novamova.com.ua/, **Весна** Багатомовний словник, багатомовна перекладачка та словники, підбірка програм і двійкових файлів для/з підтримкою української мови, художня, публіцистична та інша література http://vesna.org.ua/, **Лінгвістичний портал** http://www.mova.info/, **Проект англо-українського словника технічних термінів**

[1] Як приклад можна привести функціонування сайту http//www.slovari.ru, де представлені словники російської мови різних типів, або Асоціації лексикографів LINGVO (http//www.lingvo.ru), яка також переймається накопиченням та створенням електронних словників різних типів, насамперед, перекладних.

[2] Див. Широков В.А. Інформаційна теорія лексикографічних систем. – К., 1998. – 331 с.; Інтегрована лексикографічна система „Словники України", версія 1.03. – К.. 2001-2003, – CD-видання

[3] Див., наприклад, портал www.mova.info; сайт Технічного комітету стандартизації наково-технічної термінології (http://lp.edu.ua), www.slovnyk.net .

[4] Досить розлога інформація про українські словники в Інтернеті подана в статті перекладача-фрілансера Дмітрієвої М.М.(mailto: xmas@ukr.net)18 лютого 2003 року.

Проект англо-українського словника технічних термінів http://dict.linux.org.ua/, **Уроки державної мови** Українська літературна мова давно виробила власні вимовні, наголосові, граматичні, лексичні, стилістичні норми. Уміщуємо поради, рекомендації, які саме слова чи словосполучення найдоцільніше вживати, щоб передати потрібний зміст, робимо застереження про неправильне або небажане, невдале використання тих чи інших лексем у певних значеннях, у конкретному контексті. h, **СЛОВНИК. НЕТ – Тлумачний словник** Великий тлумачний словник сучасної української мови онлайн, містить понад 207 000 словникових статей та близько 18 000 фразеологізмів. h

Наші наукові інтереси зорієнтовані на інший напрям комп'ютерної лексикографії – діахронний. Питання збереження в електронному вигляді лексикографічних джерел, укладених до 30-х років XX сторіччя, саме української мови має принципове значення: XX століття було не лише періодом розвитку та вдосконалення мовознавчої науки, зокрема лексикографії, часом виходу найбільших словників української мови, але й періодом штучного коригування розвитку мови. Таким чином, словники, що були укладені в дорадянську добу України, є об'єктом не лише для етимологічних, діалектологічних, народознавчих розвідок, реконструкції мови того періоду, але й джерелом збагачення власного словникового запасу. „Староукраїнська лексикографія – органічна частина історії української літературної мови" [4: 285]. Вибір часових меж не випадковий, зумовлений історико-політичними обставинами, оскільки саме з цього часу в історії українського словникарства почалася нова, радянська епоха.

На початковому етапі роботи наповненням для лексикографічного корпусу слугуватимуть найбільші словники, укладені з початку XIX до 30-х років XX сторіччя. У процесі вибору об'єктів лексикографічного корпусу ми спиратимемося на праці сучасних дослідників історії українського словникарства, передовсім В. В. Німчука та Б. К. Галаса, а також мовознавців XIX – початку XX ст. Цінність цих досліджень – насамперед у окресленні тенденцій та здобутків українського словникарства зазначеного періоду, грунтовній хронологічній, типологічній, аналітичній характеристиці словників, що значно полегшить етап аналізу формальних даних словників.

Серед українських словників дорадянського періоду найвизначнішим, знаним словником є „Словарь української мови" у 4-х томах (1907-1909 рр.) за ред. Б. Грінченка. Саме цей словник і стане першим об'єктом лексикографічного корпусу. У подальших статтях плануємо подати результати підготовчої роботи до введення словника в корпус.

Оброблення текстів лексикографічного корпусу української мови здійснюватиметься засобами SGML (Standard Generalized Markup Language) – стандартної мови узагальненої розмітки у форматі TEI (Text Encoding Initiative) – системи кодування текстів, адаптованими до особливостей текстів українських словників на основі попередніх досліджень параметричних даних.

Література

1. Борис Галас. Ф. С. Шимкевич як лексикограф і українське словникарство (кінець XVIII – початок XX ст.). – Ужгород, 1995. – 300 с.
2. Демська-Кульчицька О. М. Основи національного корпусу української мови. – К., 2005. – 219 с.
3. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. – Донецьк: ТОВ „Юго-Восток, Лтд", 2003. – 184 с.
4. Німчук В. В. Староукраїнська лексикографія в її зв'язках з російською та білоруською. – К., 1980. – 303 с.
5. Українська мова у XX сторіччі: історія лінгвоциду: Док. і матеріали / Упоряд.: Л. Масенко та ін. – К.: Вид. дім „Києво-Могилянська акад.", 2005. – 399 с.
6. Широков В. А. Феноменологія лексикографічних систем. – К., 2004.–327 с.