Євгенія Карпіловська

Лексична картотека і корпус як інструменти лексикографічного моделювання слова

Потреба в створенні нових словників національної мови залишається гострою, доки в суспільстві функціонує сама така мова. Словник як продукт узагальнення, осмислення живої мовної практики завжди, в кращому разі, на крок від неї відстає. Уточню, що веду мову про словники сучасної мови, проте і до словників історичних, етимологічних кожна доба і кожний етап розвитку суспільства висувають свої вимоги. Словники української мови, і тлумачні, і перекладні, і термінологічні, і іншомовних слів, що з'явилися в українській лексикографії за період незалежності України, беззастережно довели потребу в ґрунтовному, але водночас виваженому, вдумливому перегляді і їхніх реєстрів, і їхньої пояснювальної частини, і їхнього ілюстративного матеріалу. Словники нового покоління, про які сьогодні стільки точиться гострих дискусій у світі, передбачають не лише інші технології укладання й іншу форму їхнього існування в комп'ютерному середовищі. Вони передусім потребують нових форматів для представлення наслідків моделювання мови, і це завдання можуть виконати лише мовознавці. Запорукою успіху в створенні таких форматів є надійна джерельна база словників. Далі подаю деякі міркування стосовно того, якою вона має бути для здійснення різноаспектного моделювання української мови в сучасних академічних словниках, що впродовж XX століття і нині в українській лексикографічній традиції прийнято вважати еталоном словника для освіченої і культурної аудиторії.

Комп'ютеризація інформаційної сфери життя суспільства, автоматизація лексикографічної і, ширше, наукової діяльності визначили на сьогодні форму такої джерельної бази словників. Вона становить зведення інформації про українську мову в електронній (оцифрованій) формі, тобто у формі, придатній для роботи з нею у комп'ютерному середовищі. Завдяки такій новій формі представлення відомостей про мову комп'ютер дає в руки дослідникові-мовознавцю й нові технології її опрацювання. Можливість зберігати в пам'яті комп'ютера великі масиви різноаспектної інформації, зокрема про слово як стрижневу одиницю української номінації й предикації, відкриває й нові можливості для маніпулювання інформацією, для її збирання з метою відповіді на певні завдання лексикографа. Зокрема, комп'ютеризація, як уже довела сучасна лексикографічна практика, якісно змінила образ лексичної

картотеки – основи для укладання словників, передусім тлумачних і, ширше, пояснювальних.

І це міркування ставить сьогодні перед українськими лексикографами перше важливе питання творення джерельної бази словників нового покоління: ***якою має бути нова електронна лексична картотека української мови за змістом, за своїм наповненням, а не лише за формою***. Якщо номінація, предикація та оцінка становлять три засадничі складники мовної діяльності суспільства, то лексична картотека виконує роль моделі передусім першого з них – номінаційного корпусу, або словникового фонду (складу, запасу) мови. Отже, лексична картотека на сьогодні не лише не втрачає свого значення для лексикографа, а, навпаки, її значення зростає у зв'язку з як ніколи раніше інтенсивним розвитком і оновленням української мови, можливістю оперативно відстежувати й фіксувати зміни в ній і водночас узагальнювати їх, професійно осмислювати. Сьогодні лексична картотека дедалі більше із зведення ілюстрацій уживання слова перетворюється на зведення різноаспектних відомостей про системні та текстові характеристики слова, тобто картотека ілюстрацій набуває статусу картотеки «портретів» слів. Ідею створення саме такої картотеки підтримує і наш колектив. Більше того, ми запропонували своє бачення картотеки такого типу для нової української лексики у форматі параметричної лексичної бази даних і в нашому ідеографічному словнику нової лексики «Активні ресурси сучасної української номінації» (К., 2013) подали формати «портретів» для різних типів лексичних інновацій: актуалізованої лексики, новотворів, неосемантизмів і новозапозичень.

На мою думку, така картотека сьогодні особливо актуальна саме для розв'язання проблеми оновлення корпусів українських словників: і їхніх реєстрів, і їхніх пояснювальних частин (різних форматів для різних типів пояснювальних словників: тлумачних загальномовних, термінологічних, словників іншомовних слів, тлумачно-словотвірних, ідеографічних тощо), і їхніх ілюстративних додатків, адже моделювання саме такої мобільної ділянки українського лексикону в сучасних словниках становить нині основну проблему для лексикографів. Для лексикографічного моделювання нової й оновленої, актуалізованої лексики недостатньо прикладів її уживання в контексті, ілюстрацій. Для адекватного подання форми, семантики й функціональних властивостей інновацій картотека має містити якомога вичерпнішу інформацію про весь спектр системних і текстових характеристик слова, про його місце у мовній свідомості сучасного носія української мови.

У рамах відомого трикутника відношень мовного знака Ч. Морріса для роботи лексикографа картотека має подавати відомості про слово у

площинах: 1) семантики (=відношення слова до дійсності, об'єкта позначення), 2) синтактики (=відношення слова до інших знаків у мові як знаковій системі) та 3) прагматики (=відношення мовця до слова, його ставлення до нього як до складника знакової системи і означення поняття). У форматі «портрета» слова в новій лексичній картотеці кожній з таких площин-іпостасей слова мала б відповідати окрема зона (поле чи поля) інформації залежно від спектру її типів та їхньої складності (глибини структурування). Наприклад, у форматах нашого вищезгаданого ідеографічного словника нової лексики «Активні ресурси сучасної української номінації» для моделювання «портрета» інновацій запропоновано 7 зон інформації: 1) ***зона самої інновації*** – об'єкта опису в картотеці із зазначенням можливих варіантів її написання чи наголошування, з визначенням типу такої інновації (актуалізоване слово, новозапозичення, новотвір чи неосемантизм) та хронології її побутування в українській мові за матеріалами словників і текстів; 2) ***зона відмінювання слова*** (за «Граматичним словником української літературної мови. Словозміна» (К., 2011), який уклав колектив співробітників нашого відділу на чолі з В. І. Критською (відповідальний редактор словника – Н. Ф. Клименко). У цій зоні для слів, зареєстрованих у словниках до 1991 р., подано зразок відмінювання (код словозмінної парадигми) або зазначено, що вони невідмінювані, а для нової лексики за зразками, уміщеними в словнику, подано парадигму їхнього відмінювання з можливими варіантами; 3) і 4) ***зони дефініцій слова***, відповідно, нової і старої (поданої у словниках до 1991 р.) для унаявнення нових значень актуалізованих слів або так званих неосемантизмів, нових значень уже відомих слів на зразок *вертикаль, вимір, поле, простір, світ, формат, брудний, чистий*; 5) ***зона синтагматики (сполучуваності) слова***, яка засвідчує предикацію, текстові властивості слова; у цій зоні з різним ступенем деталізації можна фіксувати типи сполук з описуваним словом, підкреслюючи передусім нові його сполучувальні властивості, зокрема внаслідок міжчастиномовних переходів (конверсії) або неосемантизації; 6) ***зона парадигматики слова***, його системні властивості, відношення з іншими словами: синонімічні, антонімічні (градуйовані мезонімами, пор. у новому значенні 'законний' тріаду прикметників *білий – сірий – чорний* (зарплата, ринок, товари) або опозиції на зразок *науковий – білянауковий, навколонауковий, квазінауковий, лженауковий, псевдонауковий – ненауковий, антинауковий*), омонімічні (пор. інновацію *альтернативник* у значеннях 'особа, яка перебуває на альтернативній військовій службі', 'особа з альтернативною, нетрадиційною сексуальною орієнтацією' і 'представник альтернативних галузей наукового знання або мистецтва (історії, літератури, музики, кінематографу)', гіперо-гіпонімічні (родо-видові, категорійно-розрядні)

відношення, наприклад, між словами *мережа* та *І(і)нтернет* в їх новому значенні ‘об’єднання комп’ютерів’; 7) ***зона епідигматики (словотворення) слова***, його словотвірного потенціалу, який охоплює всі способи морфологічного й семантичного словотворення, реалізовані таким словом, у тім числі, каламбурне словотворення, різноманітна гра з ним у мовній практиці. Остання зона дуже важлива для української мови, оскільки словотворення втримує свої позиції провідного способу її номінації і захищає її основну типологічну рису – синтетизм, однослівність.

Такий різноаспектний «портрет» слова, звичайно ж, варто сформувати не лише для нової або оновлюваної лексики, а й для лексики, стабільно відтворюваної в українських текстах різних функціональних стилів, сталої в складі українського лексикону. Таке стабільне ядро українського номінаційного корпусу, функціонуючи в мовній діяльності сучасного суспільства, також неминуче, з одного боку, виявляє рухомість, змінюваність своїх властивостей, наприклад, у сфері словозміни й словотворення, а з другого, укладання нових словників неможливе без подання такого ядра, оскільки вони мають подати в цілому реальну картину функціонування мови: як її стабільного ядра, так і периферії й мобільної перехідної зони між ними.

Безперечно, що лексикограф не може зосередити свою увагу лише на мобільних діялнках лексикону, на мовних інноваціях. У цьому переконують і цифри, які вказують на співвідношення уже відомого обсягу загальномовного лексикону з масивами нової лексики, раніше не засвідченої у загальномовних словниках. Так, обсяги реєстрів словників нової лексики, як засвідчує досвід не лише української, а й у цілому слов’янської неології, не перевищують 6 тис. слів. Наприклад, наш уже згадуваний вище словник нової лексики вмістив близько 4 тис. нових слів, що з’явилися в українській мові за останні 22 роки. Відомий «Тлумачний словник російської мови кінця XX ст.: Мовні зміни» за ред. Г. М. Скляревської (СПб, 2000), описав 5,5 тис. слів і висловів, що поповнили склад російської мови за 12 років (1985-1997), а реєстр «Словника модних слів» Вл. Новікова, нещодавно виданого в серії «Словники для інтелектуальних гурманів» (М., 2012), налічує всього 138 слів.

Реєстри нових термінологічних словників, які представляють нові сфери суспільного життя, зокрема, сферу інтернет-комунікації, наприклад, «Англо-український тлумачний словник з обчислювальної техніки, Інтернету і програмування» Е. М. Пройдакова і Л. А. Теплицького за ред. О. Л. Перевозчикової (К., 2006), уміщують 10-15 тис. нових термінів (слів і словосполук). Отже, відсоткове відношення реєстрів словників нової лексики до реєстрів загальномовних словників показує, що нові слова становлять не вище 10 % уже відомого словникового фонду (пор. з обсягом реєстру

СУМ-11 у близько 135 тис. слів), до того ж переважно це оказіональна або детермінологізована лексика, яка ще потребує випробування часом й аргументів доцільності її введення до реєстрів словників активного типу. А саме їх нам зараз бракує – словників однотомних на зразок виданого 2012 р. тлумачного «Словника української мови» за ред. В. В. Жайворонка, серії однотомних словників польської мови, наприклад, «Малого словника польської мови» за ред. Е. Соболь (1968 і 1993 років) або однотомного «Словника російської мови» С. І. Ожегова (1949, з 1992 – «Тлумачний словник російської мови» разом з Н. Ю. Шведовою) чи чотиритомного «Словника російської мови» за ред. А. П. Євгеньєвої (1957-1961, 1981-1984), так званого МАС (малий академічний словник). Невідкладне завдання сучасної української лексикографії полягає в тому, щоб на основі сучасної лексичної картотеки створити якісні, ретельно вивірені за різнофункціональними текстами реєстри саме таких активних словників, обсяги яких не перевищували б 80-90 тис. слів. Адже те, що загальну мовну діяльність суспільства забезпечують саме такі активні словники, словники так званого середнього типу, з реєстрами обсягом у 50-100 тис. слів, є аксіомою і, на моє глибоке переконання, саме на їх створенні мають зосередити свою увагу не лише лексикографи, а й усі україністи, які своїми дослідженнями мови готують ресурси для таких словників, виробляють засади лексикографічного моделювання в них слова й інших мовних одиниць, виокремлюваних зі слова (морфемних і словотвірних) чи створюваних за його участю (синтаксичних, зокрема фразеологічних). Лише підготовка таких словників нового покоління відкриває шлях до укладання нових великих тлумачних словників, словників-тезаурусів мови, нащадків гордості української лексикографії XX століття – «Словника української мови» в 11-ти томах.

Доцільність окремого підрозділу загальної лексичної картотеки, у якому б накопичувалася інформація про зміни в словниковому складі мови і у мові в цілому, бачу ще й у тому, що такий модуль картотеки став би «полігоном» для підготовки ресурсу укладання нових словників. Відстежування функціонування нової лексики в текстах різних функціональних стилів, визначення ступеня її усталеності закладає підґрунтя для прийняття виваженого рішення про введення нової лексики до загальномовного словникового фонду. Для цього лексикографи виробили вже низку кількісних та статистичних показників. Так, Н. З. Котелова запропонувала до словників-довідників нової лексики, які мають описувати слова, вже певною мірою усталені у мовній практиці протягом не менше ніж десятиліття, вводити такі неологізми, які стабільно функціонують у текстах протягом 3-4 років, а для контролю обстежувати тексти перших і останніх двох років (6: 5). Така методика, утім, була

розрахована на ручне опрацювання текстів. Сучасні комп'ютерні технології уможливлюють щорічне оброблення великих масивів текстів.

Міру коливання активності нової лексики протягом десятиліття як загальноприйнятого періоду формування нового покоління мовців спробували статистично змоделювати німецькі лексикографи С. Лібольд, Н. Тауберт і Т. Вольф – укладачі «Словника німецьких неологізмів» за ред. У. Квастхофа (8: 7-8). Вони обстежили функціонування нових слів і відомих слів з новими значеннями в газетних текстах протягом 1996-2006 рр. Для кожного року формувалися самостійні річні корпуси текстів обсягом не менше 2 млн слів і на них для кожного неологізму обчислювалася абсолютна частота вживання. Для зіставлюваності статистичних показників до корпусів відбиралися тексти тих самих газет і тієї ж тематики, тобто в такий спосіб забезпечувалася однорідність річних вибірок. У статтях словника до кожного слова додано гістограму, або графік розподілу частоти його вживання в текстах по роках. Завдяки цьому читачі словника, передусім дослідники мови, одержують картину зростання, спаду чи стабілізації активності уживання того чи того нового слова. Наприклад, гістограма до слова ***Euro*** засвідчує його перебування на далекій периферії німецького газетного лексикону до 2002 р., коли євро було введено в грошовий обіг Німеччини як її нова державна валюта. Відтоді цей неологізм, як засвідчує гістограма, демонструє стабільність функціонування.

Утім, встановлюючи ступінь активності у мові певного неологізму, слід, крім кількісних і статистичних, брати до уваги і якісні його характеристики, а саме: його роль у структурі мовної системи і тексту, його функціональний потенціал у них. Таку методику опрацювання нової лексики наш колектив реалізував у вже згаданому вище словнику «Активні ресурси сучасної української номінації». Вона дає можливість лексикографу з'ясувати роль певного неологізму у вербалізації поняття, що набуває значущості в свідомості мовця, дає змогу побачити його не ізольовано, а в складі цілісного понятєвого поля мови.

Сьогодні мовознавці, передусім лексикографи, дедалі більше уваги приділяють так званому інтегральному моделюванню мови. Це не лише передбачає тісний зв'язок словника і граматики (1), а й вивчення функціонального потенціалу слова в тексті і в системі мови (3-4). У зв'язку з цим особливого значення набуває творення корпусів мови та їх використання в укладанні словників нового покоління. Нині україністи нарешті одержали у вільному доступі корпус українських текстів, який створив колектив працівників лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка на чолі з д. ф. н. Н. П. Дарчук (5). В основу морфологічної та синтаксичної розмітки (анотування) текстів цього корпусу покладено системи автоматичного морфологічного та

синтаксичного аналізу українського тексту, створені у відділі структурно-математичної лінгвістики Інституту української мови НАН України під керівництвом д. ф. н., проф. В. С. Перебийніс (з 1962 до 2011 р. цей відділ працював у складі Інституту мовознавства ім. О. О. Потебні НАН України). Корпус налічує нині близько 50 млн. слововживань у текстах художнього, публіцистичного, наукового й офіційно-ділового стилів, має широкі хронологічні рамки (друга половина XX – початок XXI ст.) і глибоко структурований інтерфейс для пошуку інформації про слово в текстах (2).

Однак корпус можна розглядати як певну внормовану, збалансовану, кодифіковану модель мовної діяльності суспільства. Він аж ніяк не може претендувати на повноту подання її картини, яким би потужним не був (сьогодні, наприклад, Британський національний корпус (7) налічує понад мільярд слововживань). Більше того, інтерфейс корпусу унаочнює вже наявні знання про мову і не налаштований на виявлення змін у ній.

З огляду на це, крім корпусу, лексикограф має спиратися на якомога ширшу й розмаїтішу текстову базу, як в електронній, так і в традиційній паперовій формі, доки не розв'язано проблему оцифрування вже наявних текстів. Ми сьогодні маємо різноманітні електронні бібліотеки українських текстів (художніх, газетно-журнальних, наукових, офіційно-ділових, навчально-освітніх), проте 1) їхня якість і адекватність паперовим оригіналам (у разі наявності) потребують перевірки і 2) такі бібліотеки далеко не повні і не збалансовані за спектром стилів мови, показовістю їх представлення, типами текстів, їхньою хронологією.

Отже, картотека і корпус становлять модель наших знань про, відповідно, номінацію і предикацію у мові, про мовну систему і мовну діяльність спільноти. Для лексикографа вони являють собою «сполучені посудини», оскільки картотека в обговореному вище форматі «портрета слова» дає образ шуканої в тексті інформації про нього, а корпус як збалансована й кодифікована модель текстів дає уявлення про спектр уживання слова, характер його функціонування у тій чи тій сфері мовної діяльності суспільства або певного мовного осередка чи й окремого мовця. Водночас конкретна лексична картотека (картотека для створення певного словника) і корпус перебувають у ширшому просторі відомостей про мову: конкретна картотека є варіантом фундаментальної, зведеної картотеки-скарбниці мови, а корпус становить вибірку за певними критеріями з усієї сукупності наявних на сьогодні українських текстів. Взаємодія картотеки і корпусу, їх взаєможивлення є запорукою створення реальної картини складу і функціонування української мови в словниках нового покоління.

Література

1. Апресян Ю. Д. Некоторые следствия из единой модели языка для машинных фондов // Машинный фонд русского языка: идеи и суждения. – М.: Наука, 1986. – С. 90-108.

2. Дарчук Н. П. Комп'ютерне анотування українського тексту: результати і перспективи. – К.: Освіта України, 2013. – 543 с.

3. Карпіловська Є. А. Передмова // Карпіловська Є. А., Кислюк Л. П., Клименко Н. Ф., Критська В. І., Пуздирєва Т. К., Романюк Ю. В. Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики. – К.: ТОВ «КММ», 2013. – С. 6-17.

4. Карпіловська Є. А. Функціональний потенціал конкурентних моделей словотворення: параметри стабільності похідних // Функцыянальныя аспекты словаўтварэння: Дакл. IX Міжнар. навук. канф. Камісіі па славянскаму словаўтварэнню пры Мыжнар. камітэце славістаў (Мінск, 9-14 кастрычніка 2006 г.). – Мінск: ВТАА «Права і эканоміка», 2006. – С. 92-101.

5. Корпус текстів української мови – www.mova.info

6. Котелова, Н. З. Предисловие // Новые слова и значения. Словарь-справочник по материалам прессы и литературы 70-х годов; Под ред. Н. З. Котеловой. – М.: Рус. язык, 1984. – С. 3–12.

7. **BNC** – British National Corpus – www.natcorp.ox.ac.uk

8. Deutsches Neologismenwörterbuch. Neue Wörter und Wortbedeutungen in der Gegenwartssprache / Hrsg. von U. Quasthoff unter Mitarbeit von S. Liebold, N. Taubert, T. Wolf. – Berlin, New York: Walter de Gruyter, 2007. – 699 S.

Ievgeniia KARPILOVSKA

Lexical Card Index and Corpus as a Tools of Lexicographical Modelling of Word

The conception of modern multiparametrical lexical card index – the so-called card index "portraits of words" – unlike of traditional card index of quotations is concidered. Such card index of new type should include information about the system and text properties of words that would fully reflect their semantics, syntagmatics, paradigmatics, epidygmatics, chronology, pragmatics. It should interact with the language corpus and broad base of texts for the modelling of real state of modern lexicon.

Keywords: modelling of language, lexicon, lexical card index, corpus, card index "portraits of words"

УДК 811.161.2'373.43