Олена Сірук
Київський національний університет імені Тараса Шевченка

ЗАГАЛЬНОМОВНИЙ І СПЕЦІАЛІЗОВАНИЙ ТЕЗАУРУСИ: ПОРІВНЯЛЬНИЙ АНАЛІЗ

Сьогодні одним із найважливіших завдань лексикографії є проектування таких словників, які б на рівні світових стандартів задовольняли потребу сучасної інформатизованої спільноти в систематизованій лінгвістичній інформації. З огляду на це **тезауруси** як словники, які не лише фіксують, а й систематизують лексичні одиниці в межах потрібної мовної підсистеми, потрапляють у поле підвищеної уваги фахівців. Рівень розвитку інформаційних технологій в Україні дозволяє, а потреби користувача вимагають зосередитися на розробленні саме **комп'ютерних тезаурусів** різних типів: як загальномовних, так і вузькогалузевих термінологічних тезаурусів. Ця робота проводиться співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка в рамках наукових студій з формалізації мовних досліджень [2, с. 3–10; 4, с. 84–87].

Через відсутність українських ідеографічних словників (не тільки комп'ютерних, а й паперових), а також у зв'язку зі станом лексикографічних досліджень з тезаурусотворення терміносистема такої порівняно "молодої" мовознавчої галузі, як комп'ютерна ідеографія, досі практично не розроблялася. Тому в процесі роботи над укладанням "Комп'ютерного тезауруса української мови" як словника, який хоча б частково задовольнив потребу української лексикографії в загальномовних комп'ютерних тезаурусах, виникла необхідність чітко визначити і систематизувати терміни цієї лінгвістичної ділянки. Для роботи над обома проектами було здійснено огляд вітчизняної та зарубіжної літератури з лінгвістичної семантики та лексикографії, а також проаналізовано та систематизовано вже готові лінгвістичні продукти, що є в бібліотеках та в мережі Інтернет (15 паперових і понад 50 комп'ютерних словників тезаурусного типу).

"Комп'ютерний тезаурус дієслів української мови" та **"Спеціалізований тезаурус з комп'ютерної ідеографії"** – це зразки комп'ютерних ідеографічних словників одного типу, але різних підтипів за тематичною спрямованістю: загальномовного та спеціалізованого тезаурусів. Тематична специфіка ідеографічного словника визначає його **склад, структуру, особливості укладання та застосування**.

Загальномовні тезауруси як словники переважно неалфавітного типу, у яких експліцитно відображені системні семантичні відношення між одиницями, представляють лексичний склад усієї мови і, як правило, мають значний обсяг (наприклад, комп'ютерний "Тезаурус Роже", "Merriam–Webster Online Thesaurus", "Visual Thesaurus", CARMEN, SWD, EuroWordNet, BalkaNet, RussNet тощо). Загальномовний тезаурус, який включає тисячі слів і словосполучень, належать до розряду коротких. **Спеціалізовані словники**, або словники підмов, представляють терміносистему певної галузі науки. До них можна зарахувати такі комп'ютерні системи, як тезаурус НАСА з аеронавтики ("NASA Thesaurus"), сільськогосподарський тезаурус AGROVOC, тезаурус археологічних об'єктів "Archaeological Objects Thesaurus", тезаурус з астрономії "The Astronomy Thesaurus", тезаурус з біоетики "Bioethics Thesaurus", кембріджський тезаурус природничих наук "Cambridge Life Sciences Thesaurus", тезаурус біології тварин "Tesauro ICYT de Biologìa animal (CINDOC)", тезаурус інформації та документації "Thesaurus INFODATA (Thesaurus für den Bereich der Information und Dokumentation)", тезаурус з демографії "POPIN Thesaurus (Population Multilingual Thesaurus)", словник і тезаурус військової термінології "CALL Dictionary and Thesaurus (US Government)", "Тезаурус термінології гендерних досліджень" А. Денисової тощо. В Інтернеті такі термінологічні ресурси реалізовані у вигляді словника понять і термінів зі зв'язками між ними. Основне призначення словника такого типу – допомога у процесі інформаційного пошуку: на основі зв'язків тезауруса відбувається розширення запиту, а навігація за зв'язками в тезаурусі допомагає чіткіше сформулювати сам запит. Спеціалізований тезаурус, що містить 150–200 одиниць, вважається повним.

Одиниці тезаурусів. "Спеціалізований тезаурус з комп'ютерної ідеографії" нараховує 75 термінів. Для порівняння: одне семантичне поле мовленнєвої діяльності в "Комп'ютерному тезаурусі дієслів української мови" містить близько двох тисяч лише дієслівних лексико-семантичних варіантів. Слід зазначити, що, хоч "Комп'ютерний тезаурус дієслів української мови" і "Спеціалізований тезаурус з комп'ютерної ідеографії" є автономними складовими більших проектів, а саме "Комп'ютерного тезауруса української мови" та "Тезауруса з прикладної лінгвістики", суттєва відмінність у кількості одиниць на користь загальномовного тезауруса залишиться або й поглибиться за рахунок збільшення реєстру словника шляхом додавання інших частин мови, зокрема іменників, яких у мові значно більше, ніж дієслів. Спеціалізо-

ваний тезаурус на значне поповнення дієсловами розраховувати не може через особливість його одиниць. Одиниці "Спеціалізованого тезауруса з комп'ютерної ідеографії" є характерними для цього типу словників – це терміни, представлені іменниками й іменниково-іменниковими чи іменниково-прикметниковими словосполученнями в різних комбінаціях (також у формі абревіатур) обсягом від двох до чотирьох слів. Переважна частина одиниць стосується тільки предметної галузі (*комп'ютерний тезаурус, розширений КТ, методика укладання КТ*), але є й спільні з іншими мовознавчими розділами (*тезаурус, ідеографічний словник* – лексикографія; *семантичне поле, сема, лексико-семантичний варіант, антонімія, гіпонімія, синонімія* – лексикологія; *база даних, лінгвістичний процесор, лінгвістичний алгоритм* – комп'ютерна лінгвістика загалом). Невелика частина термінів об'єднується не тільки видовими та родовими відношеннями, а й синонімічними. Вони стосуються здебільшого вже усталених у лінгвістичній літературі термінів, спільних з іншими розділами мовознавства (*гіпернім* та *гіперонім, ідеографічний словник* і *тезаурус, семантичне поле* і *лексико-семантичне поле, семна структура ЛСВ* і *семний набір ЛСВ, ядро семантичного поля* і *центр семантичного поля, ядерна сема, концептуальна* і *центральна сема*).

Те, що у спеціалізований тезаурус здебільшого вносяться тільки іменники (слова цієї частини мови переважають в термінології), а в загальномовному представлені слова практично всіх частин мови разом зі стійкими сполученнями слів (фразеологізмами і прислів'ями), є ще однією відмінністю між цими словниками і спонукає звернути увагу на **відмінності в лексикографуванні дієслівної та іменникової лексики.**

Оскільки у значенні дієслова переважає сигніфікативна семантика і дієслова є представниками аналітичної лексики, дієслівне значення не співвідноситься безпосередньо з предметною сферою, а висвітлює відношення між об'єктами [6, с. 51]. Ця особливість безпосередньо впливає на методику опрацювання дієслівного матеріалу, на відмінність її від роботи з іменниками. З огляду на це **для дієслів:**

1) більш прийнятна внутрішня, сигніфікативна, заснована на аналізі понять зумовленість вибору концептів;

2) більш адекватним є індуктивний підхід до впорядкування лексем;

3) суттєвого значення набувають відношення, які базуються на словотвірних типах (деривaційна гіпонімія) і валентному потенціалі (основа для міжчастиномовних зв'язків);

4) несуттєвими є відношення типу "частина-ціле", таксономія.

Як показує досвід опрацювання англійських, іспанських, німецьких, російських тезаурусів мережі Інтернет, для дієслів значно рідше порівняно з іменником знаходилося місце в різного типу тезаурусах і ще рідше – у термінологічних тезаурусах.

Основою синоптичної схеми іменників є позамовна (тобто запозичена з об'єктивної екстралінгвістичної дійсності) картина зв'язків предметів і явищ. Категоризація іменників на денотативно-ідеографічному ґрунті зумовлена самою категоріальною природою іменників, які насамперед орієнтовані на відображення предметної дійсності [1, с. 180–181]. Отже, **для іменника:**

1) характерна зовнішня, денотативна, зумовленість вибору концептів;

2) переважно застосовується дедуктивний підхід до структурування матеріалу;

3) неістотними для створення синоптичної схеми є словотвір та валентний потенціал іменника;

4) суттєвими є відношення "частина-ціле", поширена таксономія.

Лексика термінологічних тезаурусів є переважно іменною, оскільки укладачі ідеографічних словників різних мов надавали перевагу саме іменнику.

Усі ці ознаки відображаються в теоретичних засадах побудови словника, які корелюють з мікро- і макроструктурою комп'ютерного тезауруса, зокрема зумовлюють наповнення зон у його статтях. Хоч загальна структура словникової статті для іменників і дієслів є однотипною і складається з трьох головних компонентів (заголовного слова та лексем, пов'язаних із заголовним словом відношеннями міждієслівними/міжіменниковими та міжчастиномовними), але на глибшому рівні є досить суттєва різниця. Для дієслів, поряд із синонімічними, антонімічними та родо-видовими відношеннями (що притаманне й іменникам), характерні, зокрема, частотність фонетичних варіантів, розгалужена сітка словотвірних відношень, які базуються на семантиці родів дії; мережа відношень на базі дієслівного валентного потенціалу та залежність структури словникової статті від словотвірної будови дієслова (його похідності чи непохідності).

Ще однією суттєвою відмінністю між словниками є та, що в основу спеціалізованого тезауруса кладеться домінантна наукова концепція, а синоптична схема загальномовного тезауруса будується під впливом ідеологічних і світоглядних факторів. Як зазначалося вище, загалом характер лексичного матеріалу, який подається в цих словниках, принципово відмінний. Із цього випливає, що відображення

лексичної системності в спеціалізованому тезаурусі наперед визначене позамовними чинниками, терміносистемою описуваної галузі, а загальномовний тезаурус переважно сам моделює лексико-семантичну систему, відбиває мовну картину світу.

Ю. Караулов відзначає спільне у принципах побудови загальномовних тезаурусів і правилах укладання спеціалізованих інформаційно-пошукових тезаурусів [3]. Обидва підтипи мають певні подібні **спільні** й **об'єднавчі риси**:

1) в обох словниках більш чи менш повно відображені відношення між одиницями;

2) обидва словники або мають експліцитну синоптичну схему, чи поділ універсуму на тематичні класи, або ж така схема присутня імпліцитно;

3) контекстом, чи тлумаченням, для обох словників слугує рубрика (клас умовно синонімічних слів у загальномовних тезаурусах і дескрипторна стаття в спеціальних);

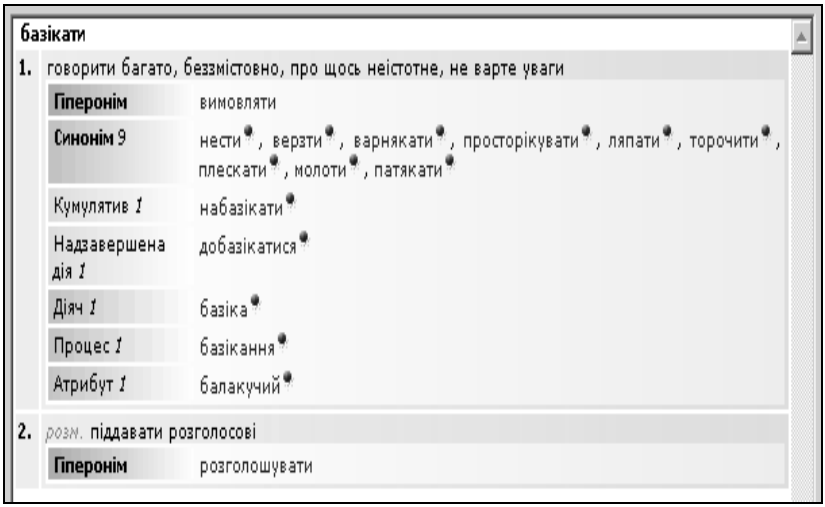
4) в обох словниках між одиницями є перехресні посилання.

Лексична семантика дієслів зумовлює відмінність ідеографічного словника іменників від аналогічного словника дієслів щодо організації його зовнішньої структури (макроструктури), у методах виявлення й опису лексичної категоризації іменників. Категоризація дієслів здійснювалася передусім на семантичній основі, при цьому застосовувався метод компонентного аналізу та ступеневої ідентифікації дієслівних значень.

Макроструктура тезаурусів. Програма "Спеціалізованого тезауруса з комп'ютерної ідеографії" має два вікна. У вікні ліворуч – пермутаційний покажчик словника у вигляді дерева термінів, рівні якого можна розгортати глибше, якщо ліворуч є позначка "+". Нульовий рівень спеціалізованого тезауруса представлено терміном *комп'ютерна лексикографія*, який є родовим для концепту першого рівня *комп'ютерна ідеографія*. Другий рівень має чотири концепти: *одиниці КТ, відношення між одиницями КТ, комп'ютерний тезаурус* та *укладання КТ*, які містять відповідно 5, 8, 10 і 6 термінів третього рівня. Максимальна глибина ієрархізації "Спеціалізованого тезауруса з комп'ютерної ідеографії" складає шість інтервалів, КТ дієслів – сім інтервалів, що відповідає умовній константі глибини будь-якого тезауруса [3, с. 186–187]. Одиниці обох тезаурусів мають тематично-алфавітне розташування.

Мікроструктура тезаурусів. Словникова стаття "Спеціалізованого тезауруса з комп'ютерної ідеографії" складається із заголовної одиниці-терміна, розміщеного у вікні ліворуч, та дефініції. Щоб знайти дефініцію терміна, його потрібно виділити мишкою і натиснути на кнопку "Тлумачення". Після цього у вікні праворуч з'явиться текст. Дефініція здебільшого складається з уточненого диференційними семами родового поняття (*Багатомовний КТ – комп'ютерний тезаурус, орієнтований на ідеографічну структуру одночасно декількох мов*), але може бути й більш розгорнутою, наближаючись до енциклопедичного визначення, коли характеризує концепт (*Комп'ютерний тезаурус (КТ) – представлений за допомогою комп'ютера ідеографічний словник. Під цим терміном об'єднуються комп'ютерна версія тезауруса та власне комп'ютерний тезаурус. КТ може бути загальномовним або спеціалізованим (за тематичною спрямованістю), одномовним чи багатомовним (за мовою виконання), мінімальним або розширеним (за повнотою викладу). Окремим видом КТ є авторський комп'ютерний тезаурус. Дослідженням КТ займається комп'ютерна ідеографія*). Семантизація заголовного слова у КТ дієслів відбувається за допомогою тлумачення з одинадцятитомного тлумачного словника української мови. Якщо дефініція – це логічне визначення поняття, встановлення його змісту та відмінних ознак, характерне для енциклопедичних та термінологічних словників, то тлумачення розкриває значення мовної одиниці з погляду наївної картини світу.

Словникова стаття КТ дієслів будується в окремому вікні. Вона може бути або тільки дієслівною (проста), або внаслідок інтеграції дієслівної частини КТ у "Комп'ютерний тезаурус української мови" розширюватися відношеннями дієслова з іменниковою, прикметниковою/дієприкметниковою та прислівниковою/дієприслівниковою лексикою (розширена). Базою виникнення таких відношень є наявність додаткових сем: 'діяч', 'інструмент дії', 'продукт дії', 'процес', 'місце, де відбувається дія', 'субстантивована дія, абстракція', 'той, що характеризується дією', 'відповідно до якостей дії'. На рис. 1 подано розширену словникову статтю дієслова *базікати*, де, окрім між дієслівних відношень, представлених гіперонімом *вимовляти*, дев'ятьма синонімами (*нести, верзти, варнякати, просторікувати, ляпати, торочити, плескати, молоти, патякати*), двома дієсловами на позначення родів дії (одним кумулятивом *набазікати* та одним дієсловом на позначення надзавершеної дії *добазікатися*), є позначені червоним тлом відношення між дієсловом та іменником (один 'діяч' *базіка* й один 'процес' *базікання*) та світло-зеленим – між дієсловом і дієприкметником (один 'атрибут' *балакучий*):



базікати

1. говорити багато, беззмістовно, про щось неістотне, не варте уваги

Гіперонім	вимовляти
Синонім 9	нести, верзти, варнякати, просторікувати, ляпати, торочити, плескати, молоти, патякати
Кумулятив *1*	набазікати
Надзавершена дія *1*	добазікатися
Діяч *1*	базіка
Процес *1*	базікання
Атрибут *1*	балакучий

2. *розм.* піддавати розголосові

Гіперонім	розголошувати

Рис. 1. **Розширена словникова стаття дієслова *базікати***

Основна **форма представлення** обох тезаурусів **– комп'ютерна.** Це бази даних у форматі Microsoft Access та програма, написана мовою C#. Для "Спеціалізованого тезауруса з комп'ютерної ідеографії" паралельно існують паперовий проект та онлайн-версія, розміщена на сторінках Лінгвістичного порталу MOVA.info у розділі "Словники" http://www.mova.info/toc.asp?PP=16&tocPath=1.

Переваги комп'ютерних тезаурусів простежуються в таких ділянках, як упорядкування матеріалу в базі даних (комп'ютерний словник є відкритою системою: базу даних можна поповнювати і редагувати, паперова версія не дає такої можливості); швидкість роботи зі словником (завдяки кільком входам, зокрема системі пошуку, а також відсиланням у дефініціях та можливості поповнювати і редагувати базу даних) та інтеграція продукту в мережу лінгвістичного програмного забезпечення (характерно тільки для комп'ютерних словників).

Система пошуку. Комп'ютерні тезауруси мають два входи: 1) за синоптичною схемою (пермутаційний покажчик) та 2) систему пошуку за лексемою та її частинами, що суттєво спрощує і пришвидшує роботу. Окрім цього, дефініції "Спеціалізованого тезауруса з комп'ютерної ідеографії" містять позначені курсивом перехресні посилання на ті терміни, які в них використовуються (**Укладання КТ** –

процес створення *КТ*, або розроблення *макроструктури КТ*, яка складається з трьох основних завдань: створення *бази даних*, *лексикографічного процесора* та вироблення формату *словникової статті*, або *мікроструктури КТ*), безвідносно до того, у межах якого з концептів вони знаходяться. У словниковій статті КТ дієслів усі лексико-семантичні варіанти, марковані відношенням до заголовного слова, є посиланнями на відповідні статті.

Застосування. "Спеціалізований тезаурус з комп'ютерної ідеографії" розрахований на філологів-фахівців та студентів філологічних спеціальностей. Може використовуватися як довідкова система та з навчальною метою. КТ дієслів має ширшу аудиторію: він може використовуватися і як багатопланова довідкова система, і як база для подальших лінгвістичних досліджень. Завдяки можливості інтеграції в мережу лінгвістичного програмного забезпечення КТ дієслів разом з пакетом додаткових утиліт були використані для аналізу особливостей авторського стилю Ліни Костенко та Василя Стуса [5, с. 246–251].

Аналіз лексикографічних матеріалів і значної кількості ідеографічних словників мережі Інтернет дав можливість 1) систематизувати термінологію з комп'ютерної ідеографії у вигляді "Спеціалізованого тезауруса з комп'ютерної ідеографії"; 2) розробити формалізовану методику укладання загальномовного "Комп'ютерного тезауруса дієслів української мови" як довідково-дослідної системи. Комплексне порівняння цих словників як прикладів загальномовного і спеціалізованого тезаурусів може знайти відображення в лекціях, спецкурсах і спецсемінарах з проблем створення комп'ютерних словників і формалізації лексичної семантики та й загалом буде корисним як для філологів, зокрема лексикографів-практиків, так і для широкого кола користувачів.

1. *Бабенко Л. Г.* Принципы категоризации именной лексики в толковом идеографическом словаре существительных русского языка // Русский язык: исторические судьбы и современность: II Международный конгресс исследователей русского языка. Труды и материалы. – М., 2004. – С. 180–181; 2. *Дарчук Н., Денисенко І., Сірук О., Сорокін В.* Ідеографічний тезаурус української мови // Вісник Черкаського університету. – Вип. 24. Серія "Філологічні науки". – Черкаси, 2001. – С. 3–10; 3. *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. – М., 1981; 4. *Сірук О.* Два підходи до побудови комп'ютерного тезауруса дієслів української мови // Українське мовознавство. – Вип. 31. – К., 2004. – С. 84–87; 5. *Сірук О.* Статистичний аналіз художнього тексту за допомогою комп'ютерного тезауруса: недоліки і переваги // Мовні і концептуальні картини Світу. – Вип. 12. – Ч. 2. – К., 2004. – С. 246–251; 6. *Уфимцева А. А.* Семантика слова // Аспекты семантических исследований. – М., 1980.