

Дарчук Н. П.
Київський національний університет
імені Тараса Шевченка

МОЖЛИВОСТІ СЕМАНТИЧНОЇ РОЗМІТКИ КОРПУСУ УКРАЇНСЬКОЇ МОВИ (КУМ)

У статті розглянуто лінгвістичні засади семантичного розмічування Корпусу української мови як четвертого етапу представлення інформації про одиниці Корпусу. В основу розмічування покладено таксономічну класифікацію корпусу російської мови, але доповнену та видозмінену. Створено програмне забезпечення для роботи в он-лайн режимі. Матеріалом слугував частотний словник публіцистичного стилю обсягом 40 тис. лексем, укладений на вибірці 16 млн словоформ українськомовного тексту.

Ключові слова: Корпус текстів, семантичне розмічування, таксономічна класифікація, таксон, тезаурус, інформаційно-пошукова система.

Комп'ютерні лінгвісти добре усвідомили, що морфологічна, синтаксична розмітка при автоматичному опрацюванні тексту може бути здійснена тією чи іншою мірою деталізації і глибини пророблення в різних мовах, але будь-які роботи зупиняються перед семантичним бар'єром. Відсутність повних описів семантики природних мов не дозволяє сподіватися на швидке здолання цього бар'єру. Настав час тривалої роботи пошуку оптимальних моделей опису семантики, їх укладання і тестування на широкому корпусному матеріалі різних методів аналізу і синтезу мовних значень.

Семантичне розмічування українськомовного тексту здійснюється в лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка і є четвертим етапом представлення інформації про одиниці Корпусу. Три попередні етапи стосуються граматичного розмічування: *Першим* етапом є морфологічне розмічування, коли кожній словоформі приписується морфологічний код частини мови і категоріальних ознак, яке дозволяє шукати приклади вживання не взагалі слів (хоча така опція присутня), а й слів у певних морфологічних формах [<http://www.MOVA.info>]. *Другим* – є синтаксичне розмічування, мета якого змодельовати синтаксичну структуру вхідного речення на рівні словосполучень і приписати інформацію про типи синтаксичних зв'язків, а також побудувати дерево залежностей речення. *Третім* – сегментування словоформ на морфи. Всі вони формалізовані, супроводжуються кількісними характеристиками – абсолютними частотами (можливими є статистичні підрахунки відносної частоти, дисперсії, коефіцієнта варіації, коефіцієнта стабільності) і залежні від морфологічної розмітки. Для морфологічного розмічування важливим є етап зняття граматичної і лексико-граматичної омонімії, який здійснюється автоматично практично на 94%, що дозволяє одержувати досить хороші результати роботи всіх етапів автоматичного опрацювання текстів Корпусу української мови. Мета розмітки – автоматично визначити граматичні параметри тексту.

Семантичне розмічування відрізняється від граматичного. Мета його – надати можливість користувачеві одержувати списки слів за заздалегідь укладеними семантичними параметрами і досліджувати їх з різних точок зору: наповненість семантичних класів, “мовна поведінка” у контексті, зсуви у значеннях як прояв системних відношень у лексиці, їхня сполучуваність тощо.

Мета статті – розкрити можливості семантичної розмітки Корпусу української мови.

Перед розробниками постало питання, за яким принципом здійснювати семантичну розмітку: за ідеографічним чи таксономічним? В основі ідеографічної класифікації лежить ієрархічний принцип – від загального до часткового, де в ролі одиниці опису є поняття. Численні ідеографічні словники, створені для різних мов, свідчать, що розробити і теоретично обґрунтувати якусь одну універсальну систематизацію не вдається. Тому нами за основу було взято таксономію Національного корпусу російської мови як апробовану вже на корпусі текстів російської мови, яка, у свою чергу, базувалася на вже працюючій із 1992 р. базі даних “Лексикограф-експерт” [5; 6] і також у процесі роботи зазнає змін. З іншого боку, не виключено, що в майбутньому всі корпуси слов’янських мов можна об’єднати в один корпус слов’янських текстів, тому бажано вже зараз формувати спільне лінгвістичне забезпечення з урахуванням лексичних особливостей національних мов.

Як зазначається в [6], створено багато лексико-семантичних класифікацій для російської мови (Кузнецова, Бабенко, Шведова), міжнародна семантична мережа WordNet; для української мови – ідеографічні класифікації [4; 10; 13], а також інтернет-ресурс UkrNet. В усіх цих лексичних класифікаціях дотримано максимально подрібнений ознаковий принцип, що аж ніяк не може задовольнити користувача Корпусу. Не можна не погодитися з К. Рахіліною [9], яка стверджує, що найкращими результатами, що можуть задовольнити користувача, є тільки ті, які ґрунтуються на лексичній базі даних із максимально жорсткою структурою і невеликою кількістю ознак (до 30). Г. І. Кустова зауважує, що для кожної частини мови розроблено свою таксономію зі своїм набором таксонів [6, с. 158]. Ураховуючи те, що користувачами Корпусу української мови є не тільки лінгвісти, а й викладачі шкіл, учні, іноземці, працівники видавництва тощо, семантична класифікація має бути зрозумілою й доступною тим, хто не має спеціальної лінгвістичної підготовки.

Передбачається, що семантично буде розмічено весь Корпус української мови. Оскільки Корпус складається з текстів різних стилів – художнього, публіцистичного, наукового, ділового – їх лексику планується розмічати по-різному. Коли йдеться про протиставлення загальної і термінологічної лексики, якою насичені науково-технічні тексти, спостерігаємо протиставлення пізнавальної глибини, науковості термінологічної лексики і наївності, побутовість загальної лексики [1; 3]. Тому, коли аналізують терміносистеми термінологи, вони розглядають її з позицій чітко визначених понять, а лексикологи – з позицій мовного вираження. До лексики наукових текстів зазвичай застосовується індуктивний підхід, суть якого полягає в моделюванні семантичних відношень у лексиці не у вигляді ієрархії (від загального до часткового), а у вигляді семантичної мережі, в якій відсутнє мотивоване розташування лексем, тобто будуються інформаційно-пошукові тезауруси для кожної науково-технічної підмови (LSP). Інформаційно-пошукові тезауруси описують певну предметну галузь і не містять інформації про загальномовну лексику. Навпаки, у публіцистичному, художньому стилях одиницею ідеографічного опису є не множина слів, а поняття, які відображають класи суспільно значущих сутностей, розрізняваних людьми, **лексем** у словнику відіграють роль **вербалізаторів понять**. Значення слова включає, крім поняттєвого змісту – сигніфікативно-денотативного компонента значення, – стилістичний, оцінний тощо. Характерним є й те, що значення слова обов’язково позначає лише дистинктивні риси об’єктів, тому у тлумачному словнику стільки різних значень, скільки слів, – поняття ж відображають глибші, істотніші семантичні властивості слів. Виходячи з цих міркувань, в Корпусі української мови

буде представлено два типи семантичної розмітки: I – **таксономічний** – для публіцистичного і художнього стилів і II – **тезаурусний** – для наукового і ділового стилів.

I – таксономічне розмічування. Ми повністю підтримуємо основні вимоги до семантичних класів у корпусній таксономічній розмітці російської мови: незалежність таксонів; базовість ознак; максимальне укрупнення класів; породження мінімального шуму на запит користувача; оптимальність результату пошуку [9, с. 226].

Лінгвістична систематика будується на перетині таксонів як багатовимірної класифікація. Таку класифікацію можна назвати логічною, оскільки вона базується на логічному принципі і виводиться апіорно. Таксони – класи, чітко розмежовані. Таксони мають як екстенціональний характер, тобто зорієнтовані на денотативний аспект лексичної семантики (напр., *назви одягу, назви рослин*), так і сигніфікативний аспект лексичної семантики (*власні назви, загальні назви*).

Лінгвістична таксономія – сукупність принципів і правил класифікації об'єктів, а також сама класифікація. Таксономія передбачає систематизацію як онтологічний результат, що відображає ієрархічну організацію. У структурі таксономії це виражається в ієрархії таксономічних категорій, пов'язаних відношенням послідовного включення від нижчого рангу до вищого. Напр., до таксону ВЛАСНІ ІМЕНА як до більш загального класу включаються:

демінутиви (*Саша, Сашко*),
імена (*Олександр*),
назви установ та підприємств (*Азовсталь*),
персонажі (*Білосніжка*),
по батькові (*Іванович*),
прізвища (*Іваненко*),
топоніми (*Київ, Оболонь, Сула*),
торгові марки (*Шанель*).

До таксону ПРЕДМЕТНІ ІМЕНА, крім іншого, входять *пристрої*, які конкретизуються таким вкладенням:

зброя (*шабля, пістолет*),
інструменти (*молоток, голка*),
меблі (*стіл, диван, шафа*),
музичні інструменти (*піаніно, скрипка, бандура*),
одяг, взуття (*капелюх, чоботи, плаття*),
посуд (*чашка, виделка*),
транспортні засоби (*автобус, сани, потяг*)

До предметних включено інформацію про мереологію (відношення “частина-ціле”, “елемент-множина”) і топологію (“поверхні”, “вмістилища”), які не є таксонами. Це дає змогу характеризувати слово за трьома параметрами, напр., *кабіна* є **пристроєм** за таксономією, **вмістилищем** за топологією і **частиною** машини (за мереологією).

Для таксономічної класифікації обрано не деревовидний, а фасетний принцип класифікації, що, з одного боку, є зручним для користувача, а з іншого, – дозволяє лінгвісту приписувати слову різні ознаки, оскільки вони часто в ньому суміщаються (див. попередній приклад), отже, і пошук здійснюватиметься за однією або комплексом ознак. Ми свідомі того, що така багатокомпонентна розмітка може видавати на запит користувачу надлишкову кількість прикладів, в яких семантичний клас заповнюватиметься словами, де ознака запиту є другорядною. І з цим треба миритися, оскільки в лінгвістичній теорії неодноразово підкреслювалося, що сприйняття лексики носіями спирається не на дискретні класифікаційні ознаки, а на

гештальти [8, S. 221]. Тільки тестування системи спрямує розробника на створення зручного запиту. Ми ж у своїй роботі намагалися уникати багатокomпонентної розмітки.

Семантична розмітка впроваджується до Корпусу української мови поетапно і в режимі on-line. Наразі опрацьовуються публіцистичні тексти, генеральна сукупність яких у Корпусі перевищує 16 млн слововживань. Укладено частотний словник цих текстів, обсягом 40 тис. різних лексем, яку розподілено за частинами мови: словник іменників, словник дієслів, словник ад'єктивів. Кожний частиномовний словник обробляється за своєю такономічною класифікацією.

При цьому зауважимо, що розмічування відбувається не за контекстом слова, а за значенням, представленим у тлумачному словнику української мови. Семантика слів у загальних рисах відображається у дефініціях тлумачного словника через ідентифікатори як основні виразники понять, вербалізаторами яких є конкретні лексеми. Якщо лексеми належать до одного семантичного класу (таксону), то в них повинні бути представлені як спільні, так і відмінні риси, зумовлені семантикою таксону, отже, глибина їх різна, але лексеми, які потрапляють в один клас, мають подібну структуру тлумачення.

При розмітці передбачено вкладені класи. Напр., на запит до Корпусу про *буттєву сферу* можна одержати набір слів, який стосуватиметься *існування* (лексеми: *життя, буття* тощо); *початку існування* (*виникнення, народження, формування, творення* тощо); *припинення існування* (*смерть* тощо).

Для автоматизованого маркування лексем словника публіцистичного стилю було створено програму, інтерфейс якої представлено на Рис. 1.

В лівій частині вікна вміщено лексеми з частиномовним кодом іменника і граматичною категорією роду (зеленим кольором позначені під'єднані до таксону лексико-семантичні варіанти – ЛСВ), у правій – таксономічна класифікація. Над нею у верхньому віконечку подається значення слова з тлумачного словника української мови в 11 томах. Проблема багатозначності при семантичному розмічуванні вирішена у такий спосіб: для кожного слова у лівій частині подається стільки значень, скільки є у тлумачному словнику української мови. Кожне зі значень (ЛСВ) розглядається як самостійне слово й автоматизовано маркується за таксономічною класифікацією.

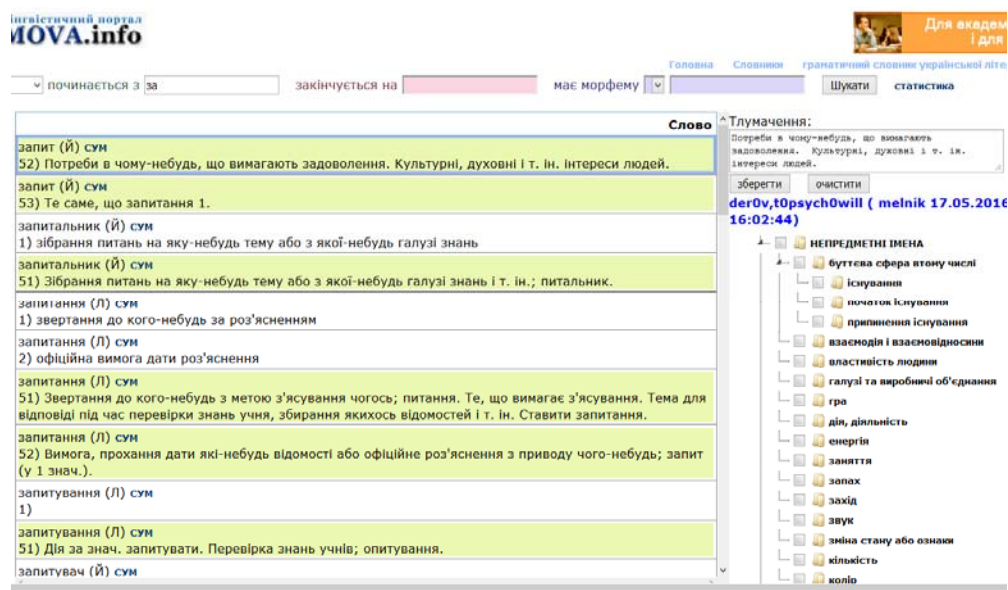


Рис. 1. Інтерфейс системи автоматизованої семантичної розмітки

Таким чином, словник іменників публіцистичного стилю обсягом 16 тис. одиниць збільшився практично у три рази. Відповідно, різні значення слова можуть належати до різних таксонів й одержують різні семантичні мітки, тобто семантичні ознаки приписуються окремо кожному ЛСВ. Ця робота здійснюється автоматизовано в режимі он-лайн. Якщо слово не має значення (неологізм, okazіоналізм, топонім тощо), йому приписується значення за контекстом та з інших джерел (напр., Вікіпедія). Це значення запам'ятовується, а потім запам'ятовується семантичний код у таксономічній класифікації.

Обов'язковим є етап редагування змісту таксономічного класу. На Рис. 2. представлено інтерфейс програми редагування, коли при натисканні миші на таксоні таксономічного дерева (ліва частина вікна) вишиковується список ЛСВ (права частина вікна) із тлумаченням до кожного ЛСВ за тлумачним словником української мови. Тільки після цього можна приступити до семантичної розмітки тестування всієї таксономії.

The screenshot shows the MOVA.info website interface. The header includes the logo 'MOVA.info ЛІНГВІСТИЧНИЙ ПОРТАЛ' and navigation links like 'Вихід | Мій профіль', 'Пошук', and 'Вибрати мову'. The main content area is titled 'Головна / Словники / Граматичний словник української мови (словозміна)'. On the left, a tree structure shows grammatical categories: Дієслово, Іменник, ВЛАСНІ ІМЕНА, НЕПРЕДМЕТНІ ІМЕНА, буттєва сфера в тому числі, взаємодія і взаємовідносини, вигляд, **влада**, властивість людини, живої істоти, властивість предметів, враження, галузі та виробничі об'єднання, гра, дія, діяльність, енергія, заняття, запах, захід, звання, звук. The right panel displays a list of LSVs for the selected 'влада' category:

1	абсолютизм	необмежена верховна влада в руках однієї людини	вилучити
1	автократія	необмежена верховна влада в руках однієї людини	вилучити
51	автократія	Форма правління, за якої одній особі належить необмежена верховна влада; абсолютизм.	вилучити
51	ареопаг	У Стародавніх Афінах - найвищий орган судової та політичної влади, що засідав на пагорбі Ареса - бога війни, звідки й дістав назву	вилучити
1	владі		вилучити
1	всевладдя	необмежена влада	вилучити

Рис. 2. Інтерфейс системи автоматизованого редагування таксонів

У своєму дослідженні ми дотримуємося також думки, що тезауруси, ідеографічні словники і в нашому випадку – таксономії, повинні створюватися окремо для кожної частини мови, оскільки кожна з них має свою концептуальну систему. Цікавою з цього приводу є думка, висловлена Н. Шведовою: “Для того, щоб уникнути різнобою у словнику, лексикограф повинен працювати не з алфавітним списком слів, а з певними угрупованнями в межах окремих частин мови. Тільки на цьому шляху можуть бути досягнуті єдність у розмежуванні значень, однотипність тлумачень” [11, с. 171]. Однак не виключено в майбутньому об'єднання різних частин мови, напр., за гіперлексемами (напр., *робити* – *робота* – *робітник* – *робочий* – [...]) із побудовою внутрішньої мережі відношень у спільнокоренових словах. Це видається доцільним, оскільки одне й те саме поняття, виражене різними частинами мови, погіршуватиме пошук при автоматичному семантичному аналізі, якщо він вестиметься тільки за тезаурусами різних частин мови.

Після опрацювання словника публіцистичних текстів будуть автоматизовано індексуватися (розмічуватися) словники текстів художнього стилю.

Безперечно, розв'язуючи завдання із семантичного розмічування, можна одержати величезну кількість інформації про властивості слів і конструкцій, яка, з одного боку, буде корисною для уточнення номенклатури таксономії, а з іншого, дає матеріал для теоретичних висновків й узагальнень. Зауважимо, що подібна розмітка не може бути самостійним модулем автоматичного семантичного аналізу, а лише слугуватиме класифікації слів за приналежністю до того чи іншого таксону.

Оскільки Корпус української мови – не просто зібрання текстів, а й інструмент пошуку, тому в перспективі можливими будуть сполучення чотирьох модулів – морфологічного, синтаксичного, морфемного та семантичного та їхніх ознак, що приписуються одиницям Корпусу (слову / словоформі, морфемі / морфу, словосполученню, реченню). Це сприятиме оптимізації користування Корпусом, а також розв'язуванню різноманітних лінгвістичних задач. Практичне використання теоретичної моделі, якою є таксономічна класифікація лексики, стимулює розвиток лінгвістики, оскільки створюються передумови нової наукової дисципліни – експериментальної лінгвістики.

1. Граматика – семантика. Це проблеми вивчення синтаксичної та семантичної сполучуваності, граматичної семантики, сталих синтаксичних конструкцій або конструкцій з двох чи трьох елементів за заданими морфологічними чи семантичними ознаками, напр., ім. (з ознакою “особа”) у дав. відм. + прийм. до + ім. у дав. відм. (*батькам не до жартів*) при визначенні семантичної функції давального відмінка іменника “батьки” (суб’єктна; *батьки не жартують*). У таких випадках користувачеві буде надано не просто мовний матеріал ім. у дав. відм.+до+ім. род. відм., а відсортований за семантичною ознакою “особа”, напр., *“Іванові не до жартів”*.

Корпус є джерелом прикладів вживання слів у конкретних текстах, що є важливим для лінгвістичного дослідження. На певний запит до Корпусу можна одержати протягом лічених секунд величезну кількість матеріалу, який можна обмежити семантичними класами:

- *мітити (цілити) в генерали* (дієсл. в інф. + прийм. в/у) + ім. наз./зн. відм. мн.;
- *взяти в служниці* (дієсл. в інф. + прийм. в/у) + ім. наз./зн. відм. мн.;
- *піти в няньки* (дієсл. в інф. + прийм. в/у) + ім. наз./зн. відм. мн.

Ця конструкція за планом вираження містить іменник у множині називного відмінка, а за змістом – у знахідному. Вручну збирати приклади у великій кількості текстів майже неможливо, а використання Корпусу полегшує це завдання, якщо при пошуку додатково включити належність до таксону “особа”. З одного боку, якщо морфологічна розмітка розширює можливості в галузі морфології, лексикографії, то семантична розмітка розширює можливості вивчення конструкцій української мови.

Для нас, розробників КУМ, важливим є дослідити можливість сполучення семантичних ознак в комбінації лексем, напр., допустимість сполучення непередметних імен і дієслів емоцій. Значну роль може зіграти семантичне розмічування у синтактико-семантичних дослідженнях, для укладання списку науково-обґрунтованих синтаксичних відношень у межах словосполучень. Напр., у словосполученнях *читати книгу, копати землю, випити чаю, помиритися з другом* встановлюються об’єктні відношення, які виражають спрямованість дії чи ознаки на предмет (*читати книгу, випити чаю, копати землю*), стосунок суб’єкта до виконуваної дії (*помиритися з другом*) тощо. Якщо задати моделі типової семантики для дієслів й іменників за таксономічним принципом і приписати членам моделі семантичні коди за

таксономічним принципом, тоді за сполучуваністю кодів можна визначити певний тип відношення. Напр.,

- читати* – дієсл. // ментальна діяльність;
- книга* – ім. // предметні імена // текст;
- копати* – дієсл. // фізичний вплив;
- земля* – ім. // предметні імена // речовина;
- випити* – дієсл. // фізіологічна сфера;
- чай* – ім. // предметні імена // їжа, напої;
- помиритися* – дієсл. // міжособистісні стосунки;
- друг* – ім. // предметні імена // особа.

Оскільки в Корпусі української мови на будь-якому тексті автоматично укладаються словники словосполучень із контекстом вживання словосполучення і здійснюється пошук слів за семантичним кодом, щоправда, в експериментальному режимі, користувач-розробник може знайти не тільки контексти, в яких уживаються дієслова певного таксономічного класу (мислення, говоріння тощо), але й перевірити сполучуваність семантичних ознак у лексемах, у зв'язку з чим семантична розмітка удосконалюється, після чого буде впроваджена у систему загальнодоступного пошуку. У перспективі планується також перевірка можливості автоматичного визначення переносних значень.

До кожної лексеми таксону добиратимуться контексти з Корпусу, на основі яких можна утворити лексико-синтаксичні фрейми як фільтри для зняття лексико-семантичної омонімії або встановлення типу синтаксичних відношень.

2. Лексика – семантика.

Ще одним важливим завданням у галузі лексикології та лексикографії є побудова семантичних словників творів певного автора. Н. Ю. Шведова вказала на надзвичайно важливу роль ідеографічного впорядкування лексики для виявлення особливостей бачення й відображення світу мовцями. Дослідниця слушно зауважила, що підмножини ЛСВ, які наповнюють нижні рівні ідеографічного дерева, не є простими групами одиниць, відібраними на підставі спільної семантики. На її думку, “такі підмножини виконують важливе власне конструктивне й інформаційне завдання. Відкриваючи перед нами певний фрагмент дійсності” [12]. У зв'язку з цим розпочато проект зі створення семантичного словника поетичних творів Лесі Українки, в якому лексика групуватиметься в межах певної частини мови за таксономічними класами.

II – тезаурусне розмічування.

У лінгвістиці відомо, що кожне повнозначне слово, зокрема іменник, може становити у тексті тему, а близькі за значенням слова в межах певного тексту утворюють тематичну домінанту тексту. Особливо це стосується науково-технічних текстів. Якщо будувати за текстом частотний словник (ЧС), то найчастотніші ЛСВ, які відображають тематичну домінанту тексту, будуть знаходитися частково у верхній його частині [2, с. 193]. Однак є значущі, проте низькочастотні ЛСВ, ігнорувати які не можна, оскільки важливими є смислові зв'язки між різними ЛСВ, в тому числі з високочастотними. Більш продуктивним шляхом укладання тематичної моделі тексту є підхід, який базується на виявленні в тексті тематично значимих ЛСВ та групування їх в лексико-тематичні групи за тезаурусним принципом. З метою виявлення змісту тексту цілком можливою є методика, яка передбачає визначення в тексті тематично значимих ЛСВ та побудову тематичної мережі на їх основі.

Вихідні положення пропонованої методики такі:

1. Тематично значимі ЛСВ доцільно групувати на основі тезаурусного дерева.
2. Групи та поля укладеного ТЗ відображають тематичне розмаїття тексту.

3. Ієрархія тем тексту визначається залежно від кількості ЛСВ, які заповнили тематичні групи та поля дерева: чим більше ЛВС потрапило до складу певної групи/поля, тим вищою є вага певної теми тексту.

Найвідповідальнішим є перше завдання методики – логіко-понятійне моделювання терміносистем, необхідне при укладанні інформаційних тезаурусів певних галузей. Це завдання здійснюється з використанням формалізованої **методики конструювання тезауруса (ТЗ)**, яка відповідає сучасним стандартам термінографії. Програма створення ТЗ і результати її роботи представлені в мережі Інтернет. Наступний крок – верифікація теоретичної тезаурусної моделі шляхом застосування її для аналізу Корпусу текстів української мови певної предметної галузі, для якої укладався ТЗ [http://www.MOVA.info.].

На першому етапі укладання ТЗ створюється інформаційно-пошукова система (ІПС) у вигляді електронної бази термінів певної предметної галузі. В алфавітному словнику для кожного слова-терміна надається тлумачення з тлумачних словників, монографій, підручників, посібників тощо.

Тезаурусний словник, крім алфавітного списку термінів і тлумачної частини до кожного терміна, містить перелік логіко-семантичних відношень між літературознавчими термінами (список запозичено з [8], але доповнено і модифіковано нами). Розроблена ІПС включає не тільки множину окремих термінів, представлених у вигляді алфавітного списку з їхніми тлумаченнями, а й самі моделі представлення зв'язків між термінами.

Побудова тезауруса (ТЗ) передбачає розкриття всіх типів відношень між термінами, основними з яких є гіпонімія (рід-вид), супідрядність на одному рівні – парціація (частина-ціле), синонімія, кореляція, асоціація, локалізація об'єкта, його призначення, функція, способи вираження функції тощо. Зміст відношень розширено настільки, щоб можна було охопити максимально широкий пласт термінів, з якими зв'язаний аналізований термін як реєстровий.

Словникова стаття побудована у вигляді анкети, “пропонованої” кожному терміну. В анкеті вміщено стандартний перелік відношень, які щодо реєстрового слова є поняттєвими. Назва відношення є двомісним предикатом **R (A, B)**, який зв'язує заголовне слово тлумачної статті (A) і введений цим предикатом термін (B) [8, с. 22].

The screenshot shows the MOVA.info website interface. At the top, there is a search bar with the text 'Термін: алгоритм' and a search button. Below the search bar, there is a list of terms and their relationships. The table below is a representation of the data shown in the screenshot.

Термін	Тип зв'язку		
іносказання	Синонім	parent	Delete
символ	Корелят	parent	Delete
металогія	Асоціація	child	Delete
метафора	Асоціація	child	Delete
парабола	Аспект	child	Delete
парабола	Параметр	child	Delete
іносказання	дивись	child	Delete
символ	Корелят	child	Delete
байка	дивись	child	Delete
байка	Інструмент	child	Delete
бернеско	дивись	child	Delete
бернеско	Інструмент	child	Delete
бестіарій	дивись	child	Delete
бестіарій	Інструмент	child	Delete
бурлеск	Інструмент	child	Delete

Рис. 3. Інтерфейс електронного словника літературознавчої термінології

Кількість семантичних відношень залежить від особливостей логіко-поняттєвих відношень предметної галузі.

Оскільки кожне з відношень має свою внутрішню службову мітку і номер в електронному форматі, це дозволяє одержувати дані, згруповані так, як це потрібно користувачеві. Представлену модель ТЗ можна розглядати як семантичний та інформаційний опис лінгвістичної термінології, яка модифікується в: а) алфавітний словник із дефініціями; б) словник синонімів; в) словник родо-видових відношень; г) автоматичний інформаційний довідник тощо.

Наступний етап – побудова тематичної моделі кожного конкретного тексту з корпусу текстів науково-технічної тематики. Методика укладання тезауруса на конкретному тексті полягає у:

- 1) лематизації та впорядкуванні слів тексту за частиномовною приналежністю;
- 2) визначенні для кожної лєми (іменникової або прикметникової) абсолютної частоти вживання;
- 3) знятті омонімії значень термінів через звертання до контексту;
- 4) побудові допоміжного реєстру слів з абсолютними частотами, які не увійшли до тезаурусу;
- 5) пошуку слів-претендентів на терміни і фіксації їх з ілюстративним контекстом у додатковому термінологічному реєстрі;
- 6) побудові тезаурусного графа термінів конкретного тексту з абсолютними частотами вживання у тексті шляхом накладання тезаурусного графа терміносистеми.

Від **семантичної мережі можна перейти і до тематичної**. З одного боку, темами є назви тематичних груп тезаурусного дерева, з іншого, – тематична мережа становить ієрархію провідних і додаткових тем. Вихідними будуть такі методичні положення: 1) тематично значимі термінологічні групи будуються на основі тезаурусного дерева; 2) ієрархія тем тексту визначається залежно від кількості термінів, які заповнили тематичні групи.

Пропонована методика дає можливість в автоматизованому режимі виявити, а потім вивчити термін у сфері функціонування.

Тезаурус ілюструє семантичну безперервність: у словнику немає і не може бути термінів, ізольованих в семантичному відношенні.

Важливість цього дослідження полягає в тому, що, по-перше, ПС у мультимедійному просторі забезпечить широке коло спеціалістів сучасним стандартизованим словником термінів, який можна поповнювати неотермінами; по-друге, здобутком є комп'ютерний інструментарій для реалізації цієї методики; по-третє, ПС сумісний з інтелектуальними системами опрацювання текстової інформації, в яких він може бути використаний як база знань та інструмент розпізнавання смислу текстів, по-четверте, створено класифікаційну модель наукового знання, що включає термінологічний словник і лінгвістичне забезпечення конкретної ПС. Якщо будуть створені за такою методикою ПС різних, напр. гуманітарних, наук, можна створювати автоматично семантичну мережу – модель плану змісту, яка є моделлю семантики галузі гуманітарного знання.

Л і т е р а т у р а :

1. *Апресян Ю. Д.* Лексическая семантика: синонимические средства языка / Юрий Дереникович Апресян. – М. : Наука, 1974. – 367 с.
2. *Баевский В. С.* Лингвистические, математические, семиотические и компьютерные модели в истории литературы / В. С. Баевский. – М. : Языки славянской культуры, 2001. – 336 с.
3. *Герд А. С.* Прикладная лингвистика / А. С. Герд. – СПб. : Изд-во СПб. ун-та., 2005. – 266, [1] с.

4. Дарчук Н. П. Комп'ютерне анотування українського тексту: результати і перспективи / Наталія Петрівна Дарчук. – К. : Освіта України, 2013. – 544 с.
5. Красильщик И. С. Предметные имена в системе “Лексикограф” / И. С. Красильщик, Е. В. Рахилина // Научно-техническая информация. – 1992. – № 9. – С. 24–31.
6. Кустова Г. И. Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы / Г. И. Кустова, О. Н. Ляшевская, Е. В. Падучева, Е. В. Рахилина // Национальный корпус русского языка: 2003–2005. – М. : Индрик. – 2005. – С. 155–174.
7. Кустова Г. И. Словарь как лексическая база данных / Г. И. Кустова, Е. В. Падучева // Вопросы языкознания. – 1994. – № 4. – С. 96–113.
8. Никитина С. Е. Тезаурус по теоретической и прикладной лингвистике (Автоматическая обработка текста) / С. Е. Никитина. – М. : Наука, 1978. – 374 с.
9. Рахилина Е. В. Задачи и принципы семантической разметки лексики в НКРЯ / Е. В. Рахилина, Г. И. Кустова, О. Н. Ляшевская, Т. И. Резникова, О. Ю. Шеманаева // Национальный корпус русского языка. Новые результаты и перспективы. – СПб. : “НЕСТОР-ИСТОРИЯ”, 2009. – С. 215–239.
10. Соколовская Ж. П. Проблемы системного описания лексической семантики / Ж. П. Соколовская. – К. : Наукова думка, 1990. – 184 с.
11. Шведова Н. Ю. О синтаксических потенциях формы слова / Н. Ю. Шведова // Вопросы языкознания, 1971. – № 4. – С. 25–33.
12. Шведова Н. Ю. Русский язык. Избранные работы / Н. Ю. Шведова. – М. : Языки славянской культуры, 2005. – 640 с.
13. Штерн І. Б. Вибрані топіки та лексикон сучасної лінгвістики: енциклопедичний словник / І. Б. Штерн. – К. : АтрЕк, 1998. – 335 с.

References:

1. Apresyan Yu. D. Leksicheskaya semantika: sinonimicheskie sredstva yazyka [Lexical Semantics: Synonymous foundations of language] / Yu. D. Apresyan. – М. : Nauka, 1974. – 367 s.
2. Baevskiy V. S. Lingvisticheskie, matematicheskie, semioticheskie i komp'yuternye modeli v istorii i teorii literatury [Linguistic, Mathematical, Semiotic and Computer Models in the History of Literature] / V. S. Baevskiy. – М. : Yazyki slavyanskoy kultury, 2001. – 336 s.
3. Gerd A. S. Prikladnaya lingvistika [Applied Linguistics] / A. S. Gerd. – SPb. : Izd-vo SPb. un-ta, 2005. – 266, [1] s.
4. Darchuk N. P. Kompiuterne anotovannia ukrainskoho tekstu: rezultaty i perspektivy [Computer Annotating of Ukrainian Text: Results and Prospects] / Nataliia Petrivna Darchuk. – К. : Osvita Ukrainy, 2013. – 544 s.
5. Krasilshchik I. S. Predmetnye imena v sisteme “Leksikograf” [Subject Names in the “Leksikograf” System] / I. S. Krasilshchik, Ye. V. Rakhilina // Nauchno-tehnicheskaya informatsiya, 1992. – No. 9. – S. 24–31.
6. Kustova G. I. Semanticheskaya razmetka leksiki v natsionalnom korpuse russkogo yazyka: printsipy, problemy, perspektivy [Semantic Annotation of Vocabulary in the Russian National Corpus: Principles, Problems, Prospects] / G. I. Kustova, O. N. Lyashevskaya, Ye. V. Paducheva, Ye. V. Rakhilina // Natsionalnyy korpus russkogo yazyka: 2003–2005. – М. : Indrik, 2005. – S. 155–174.
7. Kustova G. I. Slovar kak leksicheskaya baza dannykh [Dictionary as a Lexical Database] / G. I. Kustova, Ye. V. Paducheva // Voprosy yazykoznaniiya. – 1994. – No. 4. – S. 96–113.
8. Nikitina S. Ye. Tezaurus po teoreticheskoy i prikladnoy lingvistike (Avtomaticheskaya obrabotka teksta) [Thesaurus on Theoretical and Applied Linguistics (Automatic Text Processing)] / S. Ye. Nikitina. – М. : Nauka, 1978. – 374 s.
9. Rakhilina Ye. V. Zadachi i printsipy semanticheskoy razmetki leksiki v NKRYa [Tasks and principles of semantic marking of vocabulary in Russian National Corpus] / Ye. V. Rakhilina, G. I. Kustova, O. N. Lyashevskaya, T. I. Reznikova, O. Yu. Shemanaeva // Natsionalnyy korpus russkogo yazyka. Novye rezultaty i perspektivy. – SPb. : “NYESTOR-ISTORIYA”, 2009. – S. 215–239.
10. Sokolovskaya Zh. P. Problemy sistemnogo opisaniya leksicheskoy semantiki [Problems of the System Description of Lexical Semantics] / Zh. P. Sokolovskaya. – К. : Naukova dumka, 1990. – 184 s.
11. Shvedova N. Yu. O sintaksicheskikh potentsiyakh formy slova [On the Syntactic Potencies of the Form of a Word] / N. Yu. Shvedova // Voprosy yazykoznaniiya. – 1971. – No. 4. – S. 25–33.
12. Shvedova N. Yu. Russkiy yazyk. Izbrannye raboty [Russian Language. Selected works] / N. Yu. Shvedova. – М. : Yazyki slavyanskoy kultury, 2005. – 640 s.
13. Shtern I. B. Vybrani topiky ta leksykon suchasnoi linhvistyky: entsyklopedychnyi slovnyk [Selected Topics and Vocabulary of Modern Linguistics: Encyclopedic Dictionary] / I. B. Shtern. – К. : AtrEк, 1998. – 335 s.

Дарчук Н. П. Возможности семантической разметки корпуса украинского языка (КУЯ).

В статье рассмотрены лингвистические основы семантической разметки Корпуса украинского языка как четвертого этапа представления информации о единицах Корпуса. В основу разметки положена таксономическая классификация корпуса русского языка, но дополненная и видоизмененная. Создано программное обеспечение для работы в он-лайн режиме. Материалом послужил частотный словарь публицистического стиля объемом в 40 тыс. лексем, созданный на выборке в 16 млн словоформ украиноязычного текста.

Ключевые слова: Корпус текстов, семантическая разметка, таксономическая классификация, таксон.

Darchuk N. P. Capabilities of Semantic Tagging Within the Ukrainian Corpus.

The article views linguistic aspects of semantic tagging within the Ukrainian Corpus. The lexical content of texts of different genres, in particular, modern fiction, drama, journalism, scientific, popular scientific, and business will be provided with a specific tagging respectively. The work represents two types of tagging: I – a taxonomic one, featuring journalistic and fiction genre and II – a thesaurus-based tagging specifically for scientific and business genres.

The tagging is based on taxonomic classification applied in the Russian Corpus but extended and extra modified. There were developed the software tools for online work based on materials of frequency dictionary of journalistic style with a total volume of 40,000 lexems compiled from the sampling of 16 Million word forms of Ukrainian texts. The thesaurus-based approach is grounded on the identification of thematically relevant lexical-semantic variations and grouping them by applying a formalized method of a thesaurus construction, which meets the standards of modern terminography. There were developed the software tools for performing of two types of semantic tagging.

Keywords: linguistic corpus, semantic tagging, taxonomic classification, taxon, thesaurus, information retrieval system.

УДК 81'33(477)

Дарчук Н. П., Лангенбах М. О., Сорокін В. М., Ходаківська Я. В.
Київський національний університет
імені Тараса Шевченка

ПАРАЛЕЛЬНИЙ КОРПУС ТЕКСТІВ ПАРКУМ

Незважаючи на активний розвиток корпусної лінгвістики в Україні, досі існує велика прогалина в царині розробки паралельних корпусів. Метою роботи є формулювання основних засад творення та використання паралельного корпусу текстів ПарКУМ. Завдання, що вирішуються в ході дослідження: визначення напрямів перекладу та принципів добору текстів; вибір основних параметрів розмітки; визначення концепції роботи з матеріалом; розробка структури користувацького інтерфейсу. Подається інформація про всі типи розмітки, передбачені в корпусі: метатекстову, структурну та лінгвістичну. Окрім опису структури проекту, подано роз'яснення щодо принципів роботи з корпусом.

Ключові слова: паралельний корпус, корпусна лінгвістика, корпусна розмітка, паралельні тексти, перекладні відповідники.

Серед сучасних напрямів прикладного мовознавства чим далі, тим помітніше місце займає корпусна лінгвістика. Увага до укладання й використання лінгвістичних корпусів зумовлена, з одного боку, тим, що корпус текстів – це потужна матеріальна й інструментальна база для різноманітних наукових та практичних робіт, а з іншого, – розвитком інформаційних технологій, які суттєво спрощують процедуру створення великих колекцій лінгвістичних даних.