КОМП'ЮТЕРНА ЛІНГВІСТИКА

УДК 811.161.2'42:004

Наталія Дарчук, д-р. філол. наук., проф.,
Оксана Зубань, канд. філол. наук, доц.,
Маргарита Лангенбах, канд. філол. наук, асист.,
Ярина Ходаківська, співроб.
КНУ імені Тараса Шевченка, Київ

АГАТ-СЕМАНТИКА: СЕМАНТИЧНЕ РОЗМІЧУВАННЯ КОРПУСУ УКРАЇНСЬКОЇ МОВИ

У статті розглянуто лінгвістичні засади семантичного розмічування Корпусу української мови як четвертого етапу представлення інформації про одиниці Корпусу. В основу розмічування покладено таксономічну класифікацію Національного 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.*

Семантична розмітка текстів Корпусу української мови – це завдання нового наукового проекту, над яким працює сьогодні колектив лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка.

Семантичне розмічування текстів – четвертий етап представлення інформації про одиниці тексту у Корпусі української мови. Три попередні етапи стосуються граматичного розмічування з метою автоматичного визначення граматичних параметрів тексту. *Перший* етап, базовий для всіх наступних, – морфологічне розмічування, у межах якого кожній словоформі приписується морфологічний код частини мови і категорійних ознак, що дозволяє здійснювати пошук контекстів не тільки за заданим словом (хоча така опція також працює), а й контекстів до всіх слів за заданими морфологічними ознаками, наприклад, іменників жіночого або чоловічого роду в родовому відмінку однини, або

іменників pluralia tantum, або дієслів 1 особи множини теперішнього часу тощо. *Другий* етап – синтаксичне розмічування, мета якого змоделювати синтаксичну структуру вхідного речення на рівні словосполучень і приписати інформацію про типи синтаксичних зв'язків, а також побудувати дерево залежностей речення. *Третій* етап – сегментування словоформ на морфи. Усі етапи анотації тексту є формалізованими і дають інформацію про кількісні характеристики лінгвістичних одиниць – абсолютну частоту (можливі статистичні підрахунки відносної частоти, дисперсії, коефіцієнта варіації, коефіцієнта стабільності). Для морфологічного розмічування важливим є зняття граматичної й лексико-граматичної омонімії, яке здійснюється на 94 % автоматично, що забезпечує достовірність роботи всіх етапів автоматичного опрацювання текстів Корпусу української мови.

Семантичне розмічування відрізняється від граматичного і ставить за мету надати можливість користувачеві одержувати списки слів за заздалегідь укладеними семантичними параметрами, наприклад, таксономічними, а також досліджувати лексику за різними аспектами: наповненість таксонів, мовна поведінка в контексті, зсуви у значеннях як прояв системних відношень у лексиці тощо. У майбутньому планується також перевірка сполучуваності слів за семантичними ознаками у словосполученнях тексту та формування на цій основі списку синтаксичних відношень, а також перевірка можливості автоматичного визначення переносних значень. Ураховуючи те, що користувачами Корпусу української мови є не тільки лінгвісти, а й викладачі шкіл, учні, іноземці, працівники видавництв, семантична класифікація має бути зрозумілою й доступною для тих, 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 – їх лексику планується розмічати по-різному. Коли йдеться про протиставлення загальної й термінологічної лексики, якою насичені науково-технічні тексти, спостерігаємо протиставлення пізнавальної глибини, науковості термінологічної лексики та наївність,

побутовість загальної лексики [Апресян; Герд]. Тому, коли аналізують терміносистеми термінологи, вони розглядають її із 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, у якій відсутнє мотиване 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 суспільнозначущих сутностей, розрізнюваних людьми, лексеми в словнику відіграють роль вербалізаторів понять. Значення слова включає, крім понятєвого змісту – сигніфікативно-денотативного компонента значення, стилістичний, оцінний тощо. Характерним є й те, що значення слова обов'язково позначає лише дистинктивні риси об'єктів, тому в тлумачному словнику стільки різних значень, скільки слів, – поняття ж відображають глибші, істотніші семантичні властивості слів.

Семантичне розмічування впроваджується до Корпусу поетапно. Наразі опрацьовуються публіцистичні тексти, генеральна сукупність яких у Корпусі перевищує 16 млн. слововживань. Укладено частотний словник цих текстів обсягом 40 тис. різних лексем, які розподілено за частинами мови: словник іменників, словник дієслів, словник ад'єктивів.

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 за основу було взято таксономію Національного корпусу російсь-

кої мови як апробовану вже на корпусі текстів російської мови, яка, у свою чергу, базувалася на вже працюючій із 1992 року базі даних "Лексикограф-експерт" [Кустова; Красильщик] і також зазнала 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.

Основні вимоги до семантичних класів у корпусній таксономічній розмітці російської мови такі:

1) незалежність таксонів;
2) базовість ознак;
3) максимальне укрупнення класів;
4) породження мінімального шуму на запит користувача;
5) оптимальність результату пошуку [Рахилина : 226].

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. Наприклад, до таксону ВЛАСНІ ІМЕНА як до більш загального класу включаються:

димінутиви (*Саша, Сашко*),
імена (*Олександр*),
назви установ (*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*).

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

Створено багато лексико-семантичних класифікацій для російської мови (А. Кузнєцова, Л. Бабенко, Н. Шведова), міжнародна семантична мережа WordNet [Кустова]; для української мови – ідеографічні класифікації (Ж. Соколовська, І. Штерн, Н. Дарчук), 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, яка стверджує, що найкращими результатами, які можуть задовольнити користувача, є тільки ті, що ґрунтуються на лексичній базі даних із максимально чіткою структурою і невеликою кількістю ознак (до 30). Г. Кустова зауважує, що для кожної частини мови розроблено свою таксономію зі своїм набором таксонів [Кустова : 158].

Зупинимося на семантичному розмічуванні іменника. Воно включає три групи ознак:

1) ознаки словотвірного характеру: димінутив, аугментатив, nomen agentis, nomen femininum, віддієслівний іменник, відад'єктивний іменник;

2) предметні і непредметні іменники; власні назви;

3) власне семантичні ознаки (таксони), оцінка.

Водночас є і транскатегорійні ознаки, які діють і в зоні іменника, і в зоні дієслова, і в зоні прикметника [Рахилина : 217]. Оскільки серед іменників є девербативи і деад'єктиви, класи іменни-

ків перетинаються з дієслівними (рух, мовлення тощо) і ад'єктивними класами (колір, смак тощо).

До предметних іменників уключено інформацію про мереологію (відношення частина – ціле", елемент – множина) і топологію (поверхні, вмістилища), які не є таксонами. Це дає змогу характеризувати слово за трьома параметрами, наприклад, *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 таке багатокомпонентне розмічування може дещо розчарувати користувача, оскільки на запит будуть видаватися приклади, у яких семантичний клас заповнюватиметься словами з другорядною ознакою запиту. І з цим треба миритися, оскільки в лінгвістичній теорії неодноразово підкреслювалося, що сприйняття лексики носіями спирається не на дискретні класифікаційні ознаки, а на гештальти [Рахилина : 221]. Можливо, виходом з цього буде запит на кшталт конструкції або фрейму.

При цьому зауважимо, що розмічування відбувається не за контекстом словоформи, а за значенням, представленим у тлумачному словнику. Корпус – не просто зібрання текстів, а й інструмент пошуку, тому в перспективі можливими будуть сполучення морфологічних, синтаксичних і семантичних ознак, що сприятиме оптимізації користування Корпусом, а також розв'язуванню різноманітних лінгвістичних задач, зокрема вивченню синтаксичної та семантичної сполучуваності, граматичної семантики, сталих синтаксичних конструкцій або конструкцій із двох чи трьох елементів за заданими морфологічними чи семантичними ознаками, наприклад: іменник із семантикою 'особа' у дав. відм. + прийм. *до* + іменник у род. відм. (*батькам не до жартів*) при визначенні суб'єктної семантичної функції дав. відм. іменника *батьки*. У таких випадках

користувачеві буде надано весь контекстний матеріал корпусу, відсортований за граматичною моделлю словосполучення та семантичною ознакою ‘особа‘.

Семантика слів у загальних рисах відображається в 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, обумовлені семантикою таксону, отже, глибина їх різна, але лексеми, які потрапляють в один клас, мають подібну структуру тлумачення.

При розмічуванні передбачено вкладені класи. Наприклад, 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).

Зауважимо, що подібне розмічування не може бути самостійним модулем автоматичного семантичного аналізу, а лише слугуватиме класифікації слів за належністю до того чи того таксону.

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

мітити (*цілити*)	(дієсл. в інф. + прийм. в/у) + ім. наз./зн.
в генерали	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 врахувати належність до таксону “особа”. З одного

боку, якщо морфологічне розмічування розширює можливості в галузі морфології, лексикографії, то семантичне розмічування розширює можливості вивчення конструкцій української мови.

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

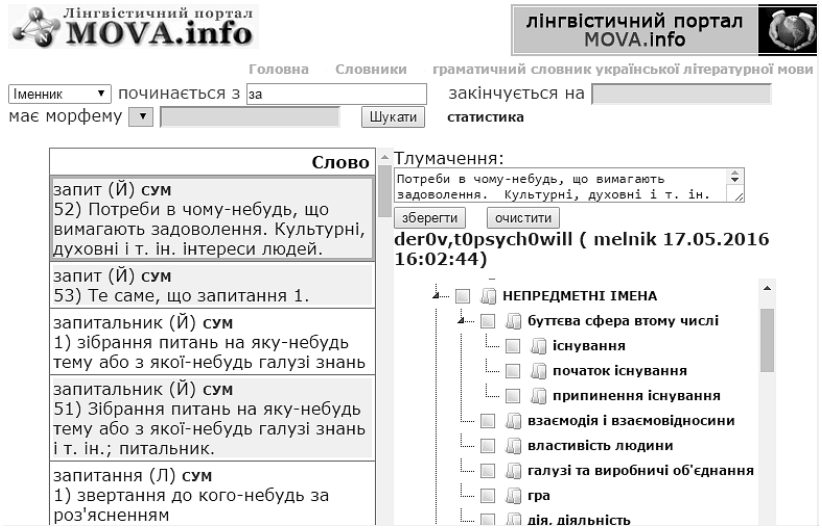

Рис. 1. Інтерфейс системи автоматизованого семантичного розмічування

У лівій частині вікна вміщено лексеми з частиномовним кодом іменника і граматичною категорією роду, у правій – таксономічну класифікацію. Над нею у верхньому віконечку подається значення слова з тлумачного словника української мови. Проблема багатозначності при семантичному розмічуванні вирішена у такий спосіб: для кожного слова у лівій частині подається стільки значень, скільки є у тлумачному словнику. Кожне зі значень (ЛСВ) розглядається як самостійне слово й автоматизовано маркується за таксономічною класифікацією. Таким чином, словник іменників публіцистичного стилю обсягом 16 тис.

одиниць збільшився практично у три рази. Відповідно різні значення слова можуть належати до різних семантичних класів і 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, оказіоналізм, топонім тощо), йому приписується значення за контекстом та з інших джерел (Вікіпедія). Це значення запам'ятовується, а потім запам'ятовується семантичний код таксономічної класифікації.

Після опрацювання словника публіцистичних текстів може постати завдання автоматизованого індексування (розмічування) словників інших (ненаукових) текстів. 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.

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.

ЛІТЕРАТУРА

1. *Апресян Ю. Д.* Лексическая семантика : синонимические средства языка / Ю. Д. Апресян. – М. : Наука, 1974. – 367 с.

2. *Герд А. С.* Прикладная лингвистика / А. С. Герд. – СПб : Изд-во С.-Петерб. ун-та. – 2005. – 266, [1] с.

3. *Дарчук Н.* Комп'ютерне анотування українського тексту: результати і перспективи / Наталія Дарчук. – К. : Освіта України, 2013. – 543 с.

4. *Красильщик И. С., Рахилина Е. В.* Предметные имена в системе "Лексикограф" / И. С. Красильщик, Е. В. Рахилина // НТИ, сер. 2. Информационные процессы и системы. – 1992. – № 9. – С. 24–31.

5. *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.

6. *Кустова Г. И., Падучева Е. В.* Словарь как лексическая база данных / Г. И. Кустова, Е. В. Падучева // Вопросы языкознания. – 1994. – № 4.

7. *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.

8. *Соколовская Ж. П.* Проблемы системного описания лексической семантики / Ж. Соколовская. – К. :Наукова думка, 1990. – 184 с.

9. *Штерн І. Б.* Вибрані топіки та лексикон сучасної лінгвістики : енцикл. слов. / І. Б. Штерн. – К. : АтрЕк, 1998. – 335 с.

REFERENCES

1. Apresjan Ju.D. (1974) Leksicheskaja semantika: sinonimicheskie sredstva jazyka [*Lexical Semantics: Synonymous Language Means*]. Moskva: Nauka, 367 p. (in Russian).

2. Gerd A.S. (2005) Prikladnaja lingvistika [*Applied Linguistics*]. Sankt-Peterburg : Sankt-Peterburg University Press, 266, [1] p. (in Russian).

3. Darchuk N. (2013) Kompiuterne anotuvannia ukrainskoho tekstu : rezultaty i perspektyvy [*Computer Annotation of Ukrainian Texts: Results and Perspectives*]. Kyiv : Osvita Ukrainy, 543 p. (in Ukrainian).

4. Krasilshhik I.S., Rahilina E.V. (1992) Predmetnye imena v sisteme "Leksikograf" [*Subject Names in the "Leksikograf" System*]. *Scientific and technical informatio, 2. series. Information processes and systems*, no 9, pp. 24-31. (in Russian).

5. Kustova G.I., Ljashevskaja O.N., Paducheva E.V., Rahilina E.V. (2005) Semanticheskaja razmetka leksiki v nacionalnom korpuse russkogo jazyka: princhipy, problemy, perspektivy [*Semantic Markup Vocabulary in the Russian National Corpus: Principe, Problems and Perspectives*]. Russian National Corpus: 2003-2005, Moskva: Indrik, pp. 155–174. (in Russian).

6. Kustova G.I., Paducheva E.V. (1994) Slovar kak leksicheskaja baza dannyh [*Dictionary as a Lexical Database*]. Problems of linguistics, no 4, pp. 96-106. (in Russian).

7. Rahilina E.V., Kustova G.I., Ljashevskaja O.N., Reznikova T.I., Shemanaeva O. Ju. (2009) Zadachi i principy semanticheskoj razmetki leksiki v NKRJa [*The Objectives and Principles of Semantic Markup Language in the Russian National Corpus*]. *Russian National Corpus. New Results and Perspectives*. Sankt-Peterburg: NESTOR-ISTORIJa, pp. 215-239. (in Russian).

8. Sokolovskaja Zh.P. (1990) Problemy sistemnogo opisanija leksicheskoj semantiki [*Problems of the Systematic Description of Lexical Semantics*]. Kiev: Naukova dumka, 184 p. (in Russian).

9. Shtern I.B. (1998) Vybrani topiky ta leksykon suchasnoi linhvistyky [*Selected topics and vocabulary of modern linguistics: encyclopedic Dictionary*]. Kyiv: AtrEk, 1998, 335 p. (in Ukrainian).

Стаття надійшла до редколегії 05.05.15

Наталия Дарчук, д-р филол. наук, проф.,
Оксана Зубань, канд. филол. наук, доц.,
Маргарита Лангенбах, канд. филол. наук, ассист.,
Ярина Ходаковская, сотрудник
КНУ имени Тараса Шевченко, Киев

АГАТ-семантика: семантическая разметка Корпуса украинского языка

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 видозмененная. 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 украинскоязычного текста.

Ключевые слова: *корпус текстов, семантическая разметка, таксономическая классификация, таксон.*

Nataliia Darchuk, Dr Hab., Professor,
Oksana Zuban, Ph D, Doc.,
Маргарита Лангенбах, Ph D, Assistant Professor,
Yaryna Khodakivska, associate
Taras Shevchenko National University of Kyiv, Kyiv

AGAT-semantics: semantic markup of the ukrainian Corpus

The article views linguistic aspects of semantic markup of the Ukrainian Corpus as the forth stage of presenting information about Corpus units. The markup is based on taxonomic classification of the Russian Corpus but with extra modification. There was developed the software tools for online work based on materials of frequency dictionary of journalistic style with a total volume of 40,000 lexemes compiled from the sampling of 16 Million word forms of Ukrainian texts.

Keywords*: linguistic corpus, semantic markup, taxonomic classification, taxon.*