АГАТ-словники як компонент автоматичного опрацювання змісту українськомовного тексту

Н. Дарчук

Інститут філології Київського національного університету імені Тараса Шевченка, Київ, Україна
Corresponding author. E-mail: nataliadarchuk@gmail.com.

Paper received 28.12.17; Revised 05.01.18; Accepted for publication 07.01.18.

https://doi.org/10.31174/SEND-PH2018-149VI42-03

Анотація. У статті розглянуто лінгвістичні засади створення електронного семантичного словника як бази для майбутнього автоматичного дослідження семантичної структури українськомовного тексту. Здійснено автоматизоване семантичне розмічування за таксономічними класами, що є частиною АГАТ-словників, за допомогою яких аналізується текст на різних рівнях мовної системи. Створена система дозволить швидко одержувати за запитом з Корпусу української мови словники лексико-семантичних груп різних частин мови із кількісними показниками. Корпус текстів і словники розміщено на мовно-інформаційному порталі www.mova.info.

Ключові слова: *АГАТ-словники, Корпус текстів української мови, таксон, семантична анотація.*

Головним компонентом комп'ютерного опрацювання тексту, його змісту є словники. Вони визначають потужність системи, за словниками ієрархізується рівневий опис мовної системи, словникова інформація визначає ефективність семантичного аналізу. Усвідомлення необхідності створення машинних словників з метою автоматичної обробки текстів дозволило укласти, наприклад, граматичний словник, словник омонімів, словник сталих сполук і фразеологічних зворотів, аналітичних конструкцій тощо. Своєрідність категоризації мовного матеріалу у цій системі так званих резидентних словників виявляється у специфічному представленні результатів морфологічного аналізу (граматичний словник), синтаксичного аналізу (словники валентностей), семантичного аналізу (потрактування і розв'язання полісемії, класифікація слів у за таксонами). Смисл цієї тенденції не тільки у тому, що в сукупності словників можуть бути відображені й якоюсь мірою відображаються всі рівні мовної структури, але й тому, що зараз у лінгвістиці пафос опису будь-якого рівня бачиться у створенні словника його елементів, відношень між ними, що і є результатом дослідницької роботи над тією чи іншою проблемою, ніби «відповіддю» на певну лінгвістичну задачу (напр., можливість укладання словника словосполучень або словника тематично пов'язаних лексем тощо). Словник виявляється найзручнішою формою узагальнення і фіксації наших знань не тільки про мову, а й про мову Шевченка, Франка, Лесі Українки, фольклору, публіцистики – сучасної і попередніх періодів. Це дає змогу отримати дослідникові крім суто лінгвістичної інформації, уяву про світоглядність автора або концептуальні характеристики стилю тощо.

Незважаючи на прикладний характер розв'язуваних завдань, при утворенні комп'ютерного словника резидентного типу слід дотримуватися обраної лінгвістичної теорії, наприклад, рівневості – розмежування морфологічних, словотвірних, лексичних, синтаксичних, семантичних, концептуальних даних. Співвідношення семантики, синтаксису й морфології у структурі словникової статті комп'ютерного словника визначається особливостями її конструювання. Кожна лексична одиниця, яка входить до бази даних, одержує опис на морфологічному, синтаксичному, семантичному рівнях у вигляді відповідного набору характеристик. Синтаксичні й морфологічні характеристики необхідні для визначення значення текстової словоформи і встановлення її синтаксичної функції у реченні. Семантичні характеристики описують значення лексичної одиниці.

Комплекс різних словників української мови в системі АГАТ призначений для автоматичного опрацювання українськомовного тексту і є інструментом аналізу Корпусу української мови, який нараховує більше ніж 80 млн слововживань (розміщений на мовно-інформаційному порталі www.mova.info). Словникова система АГАТ побудована як система словників, у якій ураховано всі типи текстових одиниць. Один зі словників, що має назву РУТА, задуманий і розвивається як самостійний словниковий комплекс багатоцільового призначення. РУТА складається з кількох частин: основну частину словника утворюють слова повнозначні та службові частини мови. Маркування лексем-термінів предметних галузей здійснюється позначками, за якими утворюються спеціальні термінологічні словники у вигляді самостійної бази даних. Неосновну частину утворюють вільні словосполучення і фразеологізми.

З лінгвістичного погляду до кожної лексеми подається різноаспектна (зональна) інформація: зона морфологічна (частина мови, для змінюваних – номер парадигматичного класу, певні граматичні характеристики для дієслів та лексико-семантичні розряди для іменників); зона синтаксична (містить припис про синтаксичну валентність і сполучуваність іменників, дієслів, ад'єктивів в українській мові); зона семантична (кожній лексемі приписується семантичний код за таксономічною класифіцією (останній розміщений поки що у тестовому режимі на мовно-інформаційному порталі www.mova.info).

Зупинимося докладніше на теоретичних засадах і прикладній реалізації семантичного словника.

Відправною точкою у міркуваннях укладачів семантичного словника АГАТ є соссюрівська теза про мову як про систему, всі елементи якої утворюють ціле. Другою відправною точкою класифікації є інша теза Соссюра: «… поза процесом мовлення слова, які мають між собою щось спільне, асоціюються в пам'яті так, що з них утворюються групи, в межах яких виявляються досить різноманітні відношення» [7, с. 121].

Основні вимоги, які автори висувають до своєї кла-

сифікації, зводяться до наступного. Семантизація слова пов'язується із розкриттям його значення. Основними компонентами значення слова можна вважати:

а) віднесеність слова до якогось відрізку позамовної дійсності;

б) його значимість (парадигматичне значення);

в) його здатність сполучуватися певним чином з іншими словами (синтагматичне значення) [5, с. 49].

Семантична розмітка словника – бази майбутнього семантичного аналізу українськомовного тексту – впроваджується поетапно. Нею буде розмічено всі лексеми частотних словників, наявних у Корпусі українськомовних текстів різних стилів: художнього, публіцистичного, ділового, крім наукового – для різних підмов створюються тезауруси із системою відношень між термінами кожної терміносистеми. Наразі опрацьовується частотний словник публіцистичних текстів, генеральна сукупність яких у Корпусі перевищує 16 млн. слововживань (розміщено на мовно-інформаційному порталі). Укладено частотний словник цих текстів, обсягом 40 тис. різних лексем, лексикон якого розподілено за частинами мови: словник іменників, словник дієслів, словник ад'єктивів, словник прислівників. Для кожної частини мови створюється своя семантична розмітка. Одиницею семантичного опису є не множина слів, а поняття, які відображають класи суспільнозначущих сутностей, розрізнюваних людьми, а **лексеми** у словнику відіграють роль **вербалізаторів понять.**

За основу було взято таксономію Національного корпусу російської мови як апробовану вже на корпусі текстів російської мови [3; 4], яка при роботі з класифікацією лексем української мови зазнала деяких змін. Ми відмовилися від ідеографічного принципу, оскільки численні ідеографічні словники, створені для різних мов, свідчать, що розробити і теоретично обґрунтувати якусь одну універсальну систематизацію не вдається. Не виключено, що в майбутньому всі корпуси слов'янських мов можна об'єднати в один корпус слов'янських текстів, бажано вже зараз формувати спільне лінгвістичне забезпечення з урахуванням лексичних особливостей національних мов.

Лінгвістична таксономія – сукупність принципів і правил класифікації об'єктів, а також сама класифікація. Таксономія передбачає систематизацію як онтологічний результат, що частково відображає ієрархічну організацію, а для кожної частини мови розроблено свою таксономію зі своїм набором таксонів. Основні вимоги до семантичних класів у корпусній таксономічній розмітці російської мови такі: незалежність таксонів; базовість ознак; максимальне укрупнення класів; породження мінімального шуму на запит користувача; оптимальність результату пошуку [6, с. 226]

В кожній понятєвій групі слова спочатку розташовуються за частинами мови (іменники, дієслова прикметники, прислівники тощо), а в межах одержаних граматично однорідних серій слів – за смисловою близькістю, що ілюструє лексичне багатство живої мови в усьому обсязі як єдине ціле. Таким же єдиним є й реальний світ, який відображається у лексиці, незважаючи на свою багатоплановість: всі події, предмети, процеси можуть розглядатися як прояв рухомої матерії, а їх взаємозв'язок підкреслюється наявністю цілого ряду об'єктивних законів руху матерії, які мають універсальний характер. Корелятом єдиного світу у пізнанні є єдине знання, закріплене у поняттях. Однією з форм фіксації понять є лексика мови [2].

Таксономічна класифікація лексики всіх частин мови складається з трьох частин: синоптичної, аналогічної та алфавітної. Для класу іменників вона описана у [1]. Дану статтю присвятимо опису синоптичної класифікації українських дієслів. **Синоптична частина** являє собою поняттєву класифікацію змістовної сторони дієслівної лексики. Таксономічна класифікація дієслів відрізняється від іменникової тим, що в ній основними одиницями аналізу є не окремі слова у певному значенні, а певні семантичні групи дієслів, що пояснюється специфікою денотативного значення цих частин мови: іменник позначає узагальнено предмет, а дієслово – тільки дію, яка відбувається з цим предметом, або її стан.

В основі організації лексичного матеріалу принцип ієрархічності спостерігаємо тільки для таксону **БУТТЄВА СФЕРА**, до якого входять: ***існування*** (*жити*); ***початок існування*** (*народитися*); ***припинення існування*** (*ліквідувати*). Умовно до буттєвої сфери можна віднести таксони: **ЯКІСНИЙ СТАН (***біліти)***; ЗМІНА СТАНУ АБО ОЗНАКИ** *(дорослішати),* які в рамках класифікації не перебувають в ієрархічних відношеннях. Решту таксонів можна охарактеризувати в такий спосіб: таксони, в яких згруповано дієслова дії та діяльності: **РУХ** *(бігти***); ПЕРЕМІЩЕННЯ ОБ'ЄКТА** *(винест)***; ПОМІЩЕННЯ ОБ'ЄКТА** *(асфальтувати)***; ФІЗИЧНИЙ ВПЛИВ на об'єкт** *(бити)***; ТВОРЕННЯ, СТВОРЕННЯ ФІЗИЧНОГО ОБ'ЄКТА** *(будувати)***; ЗНИЩЕННЯ ОБ'ЄКТА** *(спалити)***; РОЗТАШУВАННЯ ОБ'ЄКТА** *(покласти)***; МІСЦЕЗНАХОДЖЕННЯ ОБ'ЄКТА** *(лежати)***; ПОЛОЖЕННЯ ТІЛА У ПРОСТОРІ** *(валятися)***; МЕНТАЛЬНА СФЕРА** *(вірити)***; МОВЛЕННЯ** *(говорити)***; СОЦІАЛЬНА ДІЯЛЬНІСТЬ***(втручатися);* **СУСПІЛЬНО-ПОЛІТИЧНА ДІЯЛЬНІСТЬ** *(агітувати)***; ПОВЕДІНКА ЛЮДИНИ** *(вередувати)***; ФІЗІОЛОГІЧНА СФЕРА** *(гикати)***; ЗВУК***(гудіти)***; ЗАПАХ** *(духмяніти)***; СПРИЙНЯТТЯ** *(дивитися)***; СВІТЛО** *(гаснути)***.** Дієслова, які характеризують відношення: **ВІДНОШЕННЯ** *(берегти)***; МІЖОСОБИСТІСНІ ВІДНОШЕННЯ** *(аплодувати)***; СОЦІАЛЬНІ ВІДНОШЕННЯ** *(допомагати)***; ПОСЕСИВНА СФЕРА** *(мати)***; КОНТАКТ Й ОПОРА** *(торкатися)***.**

При віднесенні ЛСВ дієслова до певного таксону спостерігаємо чотири випадки:

1.) незважаючи на багатозначність дієслова, в усіх дієсловах інтегруюча сема збігається, напр., у дієслова *покачати* є п'ять ЛСВ: ЛСВ1 – «посовати, повертіти, покрутити кого-, що-небудь на чомусь або по чомусь»; ЛСВ2 – «вирівняти, вигладити качалкою та рублем (полотно, білизну і т. ін.)»; ЛСВ3 – «розкачати тісто, надаючи йому певної форми»; ЛСВ4 – «качати якийсь час помпою воду, нафту, газ і т. ін.»; ЛСВ5 – «підкидати якийсь час гуртом кого-небудь угору, виражаючи повагу, любов, захоплення і т. ін.». Хоча ідентифікатори у них різні ('посовати', 'вирівняти', 'розкачати', 'качати', 'підкидати'), ми моделюємо

денотативну ситуацію, описану в значенні, а вона пов'язана із переміщенням, тому вважаємо що такий випадок може бути інтегровано визначений таксоном ***переміщення*** і всі ЛСВ будуть зараховані до лексичної парадигми таксону ***переміщення***.

2.) У багатозначного дієслова, напр., *скриватися*, немає спільної інтегруючої семи:

ЛСВ1 – «перебувати в якому-небудь потайному місці, знаходити собі таємний притулок, ховаючись від кого-, чого-небудь, уникаючи зустрічі з кимсь»; ЛСВ2 – «таємно піти, втекти звідки-небудь, ховаючись від кого-небудь, рятуючись від чогось і т. ін.»; ЛСВ3 – «переміщаючись, віддаляючись, ставати невидимим, непомітним, зникати з поля зору»; ЛСВ4 – «не говорити, не розповідати комусь про що-небудь, приховувати щось». ЛСВ1 пов'язана із ситуацією поміщення об'єкта у щось, тому входить до лексичної парадигми таксону ***поміщення***; ЛСВ2 несе семантику руху (семи *піти, втекти*) (таксон ***рух***); ЛСВ3 – семантику переміщення (таксон ***переміщення***), а ЛСВ4 – семантику мовлення (таксон ***мовлення***).Таким чином, за всіма ЛСВ лексема *скриватися* входитиме до чотирьох лексико-семантичних парадигм.

3) Працюючи над лексичною таксономією, ми переконалися в тому, що лексика мови не допускає строгого й однозначного розбиття: одне й те саме слово у певному значенні може входити в різні лексико-семантичні групи – таксони. Наприклад, у семантиці дієслова *розширювати* – «збільшувати розміри, вставляючи що-небудь» – немає семи, яка б однозначно вказувала на відповідний таксон. У такому випадку вважаємо, що вся дефініція (слово і права частина – значення) передають денонативну ситуацію «творення, створення широкого одягу» і лексема зараховується до таксону ***творення, створення.*** Однак інфінітив дієслова «збільшувати» є деад'єктивом від прикметника *більший*, отже, в його семантиці є вказівка на якість, тобто робити щось більшим. Тому у дієслова *розширювати* буде комбінований семантичний код, який вказуватиме на віднесеність дієслова до двох таксонів – ***творення, створення*** і ***якість***. Це змусило виробити й використати систему перехресних посилань, що є насправді демонстрацією семантичної поліфонічності слова, яка реалізується в різних диференційних елементах та ознаках. Між тим така розкладена у таксони лексика є єдиною, цілісною та спаяною численними зв'язками та відношеннями.

4) До таксону входять і синонімічні, й антонімічні пари, напр., *одягти, роздягти* (таксон ***поміщення***); *зімкнути, розімкнути* – таксон ***поміщення***.

Оскільки кожне зі значень (ЛСВ) розглядається як самостійне слово й автоматизовано маркується за таксономічною класифікацією, словники іменників та дієслів публіцистичного стилю за обсягом збільшився практично втричі (загальна кількість одиниць класифікації 47554 ЛСВ). Якщо слово не має значення (неологізм, оказіоналізм, топонім тощо), йому приписується значення за контекстом та з інших джерел (Вікіпедія). Це значення запам'ятовується в базі даних при комп'ютерній обробці матеріалу, а потім запам'ятовується і семантичний код у таксономічній класифікації.

В **аналогічній частині** таксономії дієслова кожна з 37 поняттєвих груп наповнюється ЛСВ, причому поняттєві групи слідують в алфавітному порядку заголовкових слів-понять. **Алфавітна частина** є не що інше, як звичайний тлумачний словник сучасної української мови, в якому слова в кожному із своїх значень не тільки супроводжується вербальним визначенням, але й вказівкою на їх місце в таксономічній класифікації. Синоптичний словник, який дозволяє витлумачити максимальну кількість слів, є не що інше, як рівень родових слів-понять. Його ідентифікація досягається шляхом аналізу лексико-семантичних класів умовної еквівалентності.

Родовим словом-поняттям або елементом базисного словника у вигляді таксону можна визнати таке слово, виділення якого з класу забезпечується мінімальним числом диференційних ознак. Опрацьовуються всі без виключення ЛСВ і навіть ті, в яких є переносне значення. Самі лексико-семантичні класи умовної еквівалентності – таксони, які об'єднують слова із загальними інваріантними властивостями, задаються у парадигматичній частині словника і характеризуються вказівкою на їхнє місце у загальному плані класифікації спеціальним внутрішнім семантичним кодом, прихованим від користувача, але важливим для різного роду сортування і пошуку за певними параметрами.

Таку класифікацію можна назвати логічною, оскільки вона базується на логічному принципі і виводиться апріорно. Таксони – класи, чітко розмежовані. Таксони дієслова мають екстенсіональний характер, зорієнтовані на сигніфікативний аспект лексичної семантики на відміну від іменникової семантики, для якої характерний денонативний характер. Для таксономічної класифікації обрано не деревовидний, а фасетний принцип класифікації, що, з одного боку, є зручним для користувача Корпусом, а з іншого – дозволяє лінгвісту-оператору приписувати слову різні ознаки, оскільки вони часто в ньому суміщаються. При цьому зауважимо, що розмічування відбувається не за контекстом словоформи, а за значенням, представленим у кількох тлумачних словниках української мови, щоб лінгвіст мав змогу обрати найкоротше і найточніше визначення ЛСВ, оскільки передбачається на запит користувача Корпусом видавати не тільки лексеми – вербалізатори понять, але й контексти, в яких вживатиметься ЛСВ, із спливаючою підказкою про значення ЛСВ із тлумачного словника.

Зауважимо, що подібна розмітка не може бути самостійним модулем можливого автоматичного семантичного аналізу, а лише слугуватиме класифікації слів за приналежністю до того чи іншого таксону. Таксономічну класифікацію можна розглядати в аспекті зв'язку із Корпусом української мови, заради якого вона і була створена. У поєднанні з Корпусом таксономічний словник суміщатиме у собі редукований ідеографічний словник і контекстний словник слововживань. Це значить, що, з одного боку, кожне слово у ньому подаватиметься в оточенні всіх семантично близьких йому слів, а з іншого – кожне слово супроводжуватиметься описом його сполучуваності. Такий словник може бути новим типом семантичного словника. Слід наголосити на великому значенні тлумачного словника. При створенні автоматичного семан-

тичного словника української мови дефініція буде використана двічі: з одного боку, як носій характеристики означального, як семантичний еквівалент слова, з іншого — як модель синтаксичної сполучуваності аналізованого слова.

ЛІТЕРАТУРА

1. Дарчук Н., Зубань О., Лангенбах М., Ходаківська Я. АГАТ-семантика: семантичне розмічування Корпусу української мови / Українське мовознавство. - Випуск 46/1, 2016 - С. 92-103.
2. Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. / Ю.Н.Караулов. – Изд-во «Наука», 1981. – 363 с.
3. Красильщик И.С., Рахилина Е.В. Предметные имена в системе «Лексикограф» // НТИ, сер. 2 – 1992. – № 9. – С. 24 – 31.
4. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы / Национальный корпус русского языка: 2003-2005.– М.: – Изд-во «Индрик». – 2005. – С. 155-174.
5. Морковкин В.В. Идеографические словари. М. Изд-во МГУ. – 1978, с. 49.
6. Рахилина Е.В., Кустова Г.И., Ляшевская О.Н., Резникова Т.И., Шеманаева О.Ю. Задачи и принципы семантической разметки лексики в НКРЯ / Национальный корпус русскогоязыка. Новые результаты и перспективы. – Санкт-Петербург: «НЕСТОР-ИСТОРИЯ» – 2009. С.215 – 239.
7. Соссюр Ф. Курс общей лингвистики. М., Соцэгиз, 1933, с. 121

REFERENCES

1. Darchuk N.., Zuban O., Langenbach M., Khodakivska Y. AGAT-semantika: semantychne rozmitchuvannya Korpusu ukrayinskoyi movy/ Ukrayinske movoznavstvo. - Vypusk 46/1, 2016 - s. 92-103.
2. Karaulov Y.N. Lingvisticheskoye konstruirovaniye i tezaurus literaturnogo yazyka. / Y.N. Karaulov. – Izd-vo «Nauka», 1981. – 363 s.
3. Krasilshyk I.S., Rakhilina E.V. Predmetniye imena v sisteme «Lwksikograph» // NTI, ser. 2 – 1992. - № 9. – S. 24 – 31.
4. Kustova G.I., Lyashevskaya O.N., Paducheva E.V., Rakhilina E.V. Semanticheskaya razmetka leksiki v natsionalnom korpuse russkogo yazyka: printsypy, problemy, perspektivy/ Natsyonalniy korpus russkoho yazyka: 2003-2005.– M.: – Izd-vo «Indryk». – 2005. S. 155-174.
5. Morkovkin V.V. Ideographicheskiye slovari. M. Izd-vo MGU. 1978, s. 49.
6. Rakhilina E.V., Kustova G.I., Lyashevskaya O.N., Reznikova T.I., Shemaneva O.Y. Zadachi i printsypy semanticheskoy razmetki leksiki NKRY / Natsyonalnoho korpusa russkoho yazyka. Noviye zadatchi i perspektivy – Sankt-Peterburg: «NESTOR-ISTORIYA» – 2009. S.215- 239.
7. Sossur F. Kurs obshey lingvistiki. M., Sotsegiz, 1933, s. 121

AGAT-dictionaries as a component for automated Ukrainian text content management

N. Darchuk

Abstract. The article views linguistic aspects for development of electronic semantic dictionary as the basis for the future-oriented automated analysis of semantic structure of Ukrainian texts. It provides available automated semantic tagging (metadata) according to taxonomic classes which is utilized in AGAT-dictionaries and enables to analyze texts at different levels of language system. The developed system enables to display as per request from the Ukrainian Corpus dictionaries of lexical and semantic groups of different parts of speech with statistical data. The text corpus with dictionaries are available at the linguistic web-site www.mova.info.

Keywords: *AGAT-dictionaries, The Ukrainian corpus texts, taxon, semantic annotation.*

АГАТ-словари как компонент автоматической обработки смысла украинскоязычного текста

Н. Дарчук

Аннотация. В статье рассмотрены лингвистические основы создания семантического словаря как базы для будущего автоматического исследования семантической структуры украинскоязычного текста. Осуществлена автоматизированная разметка по таксономическим классам, которые являются частью АГАТ-словарей, с помощью которых анализируется текст на всех уровнях языковой системы. Созданная система позволит быстро извлекать по запросу из Корпуса текстов украинского языка словари лексико-семантических групп разных частей речи с количественными показателями. Корпус текстов и словари размещены на информационном портале www.mova.info.

Ключевые слова: *АГАТ-словари, Корпус текстов украинского языка, таксон, семантическая аннотация.*