

Тетяна ГРЯЗНУХІНА

Український мовно-інформаційний фонд НАН України
пр-т Голосіївський, 3, м. Київ, 03039, Україна
Електронна пошта: umiftatyana@gmail.com
<https://orcid.org/0000-0002-8055-442X>

Тетяна ЛЮБЧЕНКО

Український мовно-інформаційний фонд НАН України
пр-т Голосіївський, 3, м. Київ, 03039, Україна
Електронна пошта: t.lyubch@gmail.com
<https://orcid.org/0000-0002-7390-0427>

Ірина ОЛЕКСАНДРУК

Інститут мовознавства ім. О. О. Потебні НАН України
вул. Грушевського, 4, м. Київ, 01001, Україна
Електронна пошта: ira_oleksandruk@ukr.net
<https://orcid.org/0000-0003-2701-1714>

**РОЗРОБЛЕННЯ МОВНО-ІНФОРМАЦІЙНОГО
ІНСТРУМЕНТАРІЮ АВТОМАТИЗОВАНОГО
ЕКСТРАГУВАННЯ З ТЕКСТУ СЕМАНТИЧНОЇ І
КОГНІТИВНОЇ ІНФОРМАЦІЇ**

У статті йдеться про роль семантичних досліджень для створення інтелектуальних систем автоматизованого опрацювання мовної інформації. Окреслено підхід до розроблення мовно-інформаційного інструментарію автоматизованого екстрагування з тексту семантичної і когнітивної інформації. порушено питання підвищення ефективності якісного та швидкого опрацювання цифрової інформації, створення високоінтелектуальних швидкодіючих інформаційних систем.

Теоретичним підґрунтям дослідження є розроблена академіком В. А. Широким лексикографічна теорія знакових семантичних систем і побудована в її межах теорія семантичних станів одиниць природної мови, яка ґрунтується на засадах лексичної семантики і орієнтована на використання комп'ютерних технологій у семантичних дослідженнях.

Джерельною базою дослідження обрано інтегровану лексикографічну систему, що створюється в Українському мовно-інформаційному фонді НАН України, ядерним компонентом якої є комп'ютерна версія «Словника української мови» у 20 томах, який є найбільш репрезентативним джерелом семантичної інформації, що

Цитування: Грязнухіна Т., Любченко Т., Олександрук І. Розроблення мовно-інформаційного інструментарію автоматизованого екстрагування з тексту семантичної й когнітивної інформації. Мовознавство. 2023. № 3. С. 50–63. <https://doi.org/10.33190/0027-2833-330-2023-3-003>

Citation: Gryaznukhina T., Lyubchenko T., Oleksandruk I. (2023). Rozroblennia movno-informatsiinoho instrumentariu avtomatyzovanoho ekstrahuvannia z tekstu semantichnoi y kohnityvnoi informatsii [Development of language and information tools for automated extraction of semantic and cognitive information from the text]. Movoznavstvo, (3), 50–63. [In Ukrainian]. <https://doi.org/10.33190/0027-2833-330-2023-3-003>



Стаття опублікована за ліцензією CC BY-SA 4.0 (<http://creativecommons.org/licenses/by-sa/4.0/>)

забезпечує опис лексичного значення слова за багатьма параметрами із залученням прагматичного, конотативного, синтагматичного і лінгвістичного контекстів.

У дослідженні для побудови мовно-інформаційного інструментарію автоматичного розпізнання семантичної і когнітивної інформації в текстах, написаних природними мовами, застосовується формалізм теорії семантичних станів.

Запропонований семантичний класифікатор забезпечує експліцитне представлення екстрагованої з тексту інформації, що уможливило подальше програмне її опрацювання завдяки уніфікованій інтерпретації операторів формули семантичного стану і результатів семантичної індексації тексту.

Декларується доцільність застосування розробленого мовно-інформаційного інструментарію для розв'язання таких важливих завдань комп'ютерної лінгвістики, як визначення формальних критеріїв розмежування лексичних омонімів і лексико-семантичних варіантів багатозначного слова, коректне встановлення структурних протиставлень «значення — відтінки», «лексема — лексико-семантичні варіанти» в тлумаченні значення останнього, завдання уніфікації опису лексичного значення слів одного семантичного поля.

Ключові слова: автоматичне опрацювання інформації, теорія знакових лексикографічних систем, інтегрована лексикографічна система, формула семантичного стану, семантична і когнітивна інформація.

В умовах сьогодення, коли обсяги цифрової інформації неухильно зростають в усіх сферах людської діяльності — економічній, політичній, культурно-просвітницькій, науковій, — питання підвищення ефективності якісного та швидкого її опрацювання стає дедалі актуальнішим. Проблема подолання часового бар'єра людських можливостей щодо засвоєння і опрацювання інформаційної лавини особливо загострилася з появою інтернету.

Загально визнаним способом розв'язання цих проблем є впровадження комп'ютеризації в інформаційну сферу.

Через те що лівова частка інформації міститься в текстах, написаних природними мовами, увага розробників інформаційних систем зосереджена на завданнях автоматичного опрацювання саме цього виду мовлення. Ідеться про створення високоінтелектуальних швидкодіючих інформаційних систем, у функції яких входить:

- автоматичне (із залученням методів комп'ютерної лінгвістики) екстрагування інформації, записаної в тексті;

- представлення основного змісту цієї інформації у формі, придатній для подальшого застосування до неї апарату логічних операцій з метою виокремлення з аналізованого тексту знань про описуваний у ньому фрагмент навколишнього світу (певну предметну галузь);

- оцінювання ступеня інформаційної новизни отриманих знань шляхом зіставлення їх зі змістом інших документів (текстів) з тієї самої тематики;

- у випадку позитивної оцінки формування нових знань про певну предметну галузь і включення їх у сферу комунікації «людина — комп'ютер — людина».

Теоретичним *підґрунтям* можливості застосування комп'ютера і точних методів під час автоматичного опрацювання мовної інформації є такі положення сучасної лінгвістики:

– існування неоднозначного зв'язку між змістом і формою мовного знака як двосторонньої сутності, що знаходить своє вираження в єдності матеріальної форми і внутрішнього змісту (значення);

– високий ступінь системної організації лексичної підсистеми мови, яка є основним носієм інформації про властивості об'єктів позамовної дійсності, а також її багатомірність, зумовлені системним характером мови як знакової системи [Апресян; Крушевский; Леонтьева; Пиотровский];

– спроможність мови виконувати в органічній єдності свої головні функції — когнітивну (бути основним інструментом формування думки і основним знаряддям збереження та опрацювання інформації) і комутативну (бути універсальним інструментом, за допомогою якого відбувається обмін цією інформацією в процесі комунікації).

Мета статті — продемонструвати можливість і доцільність побудови на основі формалізму теорії семантичних станів мовно-інформаційного інструментарію автоматичного екстрагування й опрацювання семантичної і когнітивної інформації з текстів.

Технічний аспект здійснення автоматизації інформаційної сфери — швидкість і обсяг сучасних комп'ютерів — нині можемо вважати виконаним. Це стосується залучення комп'ютера до процесу опрацювання великих масивів інформації, розв'язання класифікаційних завдань, які вимагають великих затрат часу на виконання операцій зіставлення за багатьма параметрами в супроводі статистичного оцінювання. Саме завдяки цьому й відбувся за останні тридцять років стрімкий розвиток систем автоматичного опрацювання текстової інформації:

– від перших систем машинного перекладу з послівним підрядковим перекладом, зорієнтованим на тексти вузької тематики, до систем-перекладачів у смартфонах;

– від систем автоматичного редагування, які, по суті, здійснювали орфографічний контроль і видавали застереження щодо занадто великої довжини речень або недоречного багаторазового повторення того самого слова в межах речення, до автоматичних редакторів з розробленими модулями синтаксичного і семантичного контролю. В автоматичних редакторах передбачається не тільки виявлення помилконебезпечних місць у тексті, але й здійснення автоматичного виправлення семантико-синтаксичних помилок, що стосуються неправильних схем дієслівного й іменного керування, організації підрядного синтаксичного зв'язку як у середині речення, так і між предикативними частинами складного речення, або некоректного вибору слова на позначення певного поняття, про яке йдеться в тексті.

Для флексивних мов практично розв'язаними можна вважати також проблеми автоматичного лексико-морфологічного та морфолого-синтаксичного аналізу тексту, який є стартовим необхідним етапом для проведення будь-якого виду роботи комп'ютера з мовною інформацією [Грамматичні системи...; Любченко; Морфологический анализ...; Синтаксический анализ...; Шевченко].

У системах автоматичного опрацювання текстової інформації на результати роботи алгоритмів автоматичного морфологічного й синтаксичного аналізу спираються семантичні аналізатори [Дарчук 1983б], у функції яких безпосередньо входить здійснення автоматичного розпізнання змісту текстової інформації шляхом лінгвістичного аналізу поверхневої семантичної структури тексту на рівні лексичного значення його одиниць. У результаті роботи

цих аналізаторів отримуємо семантичне представлення (СемП) опрацьованого тексту у формі, яка забезпечуватиме експліцитне відображення референтного і реляційного компонентів лексичного значення текстових одиниць, роблячи тим самим доступною для комп'ютера інформацію, по-перше, про об'єкти навколишнього світу (референтне значення) і, по-друге, про зв'язки між ними (реляційне значення) в описуваній ситуації, що цілком відповідає поняттю «комп'ютерне розуміння інформації».

Завдання залучення комп'ютера до процесу екстрагування із семантичної інформації, зафіксованої в СемП, наукових знань про описуваний у тексті фрагмент навколишнього світу щодо іманентних властивостей об'єктів відповідної предметної галузі вирішується на перетині кількох наук — філософії, логіки, інформатики, лінгвістики, психології. У дослідженнях з інформаційних технологій загально визнаною формою представлення знань у комп'ютерних системах є бази знань, побудовані у вигляді предметних онтологій (ПО), зорієнтованих на конкретні сфери наукових досліджень предметних галузей [Рубашкин; Nirenburg, Raskin; Staab, Studer].

Структурними елементами ПО є концепти (поняття, категорії), атрибути концептів (властивості, ролі, функції), зв'язки між концептами (фрейми, концептуальні мережі). Таке структурування предметних галузей відповідає загальномовній картині світу щодо поділу його на предмети, ознаки, відношення, дії [Уемов], отриманій у результаті когнітивно-мовленнєвої діяльності людей і закріпленій у їхній свідомості.

Головна роль у формуванні конкретних ПО належить фахівцям з відповідних галузей досліджень.

Зосередження уваги на лінгвістичних аспектах когнітивного процесу як визначального фактора інтелектуальної діяльності людини відкриває шлях для здійснення автоматичного розпізнання змісту текстової інформації та її подальшого опрацювання, спрямованого на продукування нових знань через визначення кореляцій між елементами СемП і елементами ПО.

Можливість встановлення відповідності між заданою в предметній онтології концептуальною картиною дійсності та інформацією про неї з тексту, зафіксованою в СемП, зумовлена поняттєво-мовною природою концептів, яка виявляється у співвіднесеності їх з категоріями «поняття» і «сигніфікативне значення» слова. Тобто йдеться про вербалізацію результатів когнітивно-мовленнєвої діяльності людини, спрямованої на систематизацію знань про картини навколишнього світу, при якій знання фіксуються в поняттях, а ті в свою чергу відображаються в лексичному значенні слова.

Якщо розроблення концептуальної системи конкретної ПО здійснюють спеціалісти з відповідної предметної галузі та спеціалісти з інформатики, то виявлення поняттєвого змісту в лексичному значенні слів перебуває в компетенції лінгвістів-спеціалістів з комп'ютерної лексикографії. Основною вимогою до виконання завдання виявлення поняттєвих компонентів у лексичному значенні слова є представлення останніх в експліцитному вигляді, який давав би змогу, по-перше, програмно (автоматично) у межах лексичної підсистеми одного тексту здійснювати над цими компонентами операції з визначення кількісної переваги певної множини слів з однаковим поняттєвим змістом в їхніх значеннях над рештою слів. Виділені слова можуть вважатися ідентифікаторами тематичної характеристики цього тексту. По-друге, зіставлення (із застосуванням статистичних методів) різних текстів за експлікованими в значеннях їхніх слів наборами поняттєвих компо-

нентів має уможливлювати здійснення автоматичного розподілу масивів різних текстів за ознакою їхньої тематичної схожості, а всередині групи текстів однієї тематики — визначення ступеня семантичної близькості між ними. І нарешті, експліцитна форма представлення поняттєвого змісту в лексичному значенні слів має відігравати роль перехідного містка при встановленні кореляцій між структурними елементами семантичного представлення текстів і предметними онтологіями відповідних предметних галузей, а саме: забезпечувати правильне визначення місця текстової одиниці серед елементів ПО; актуалізувати в тексті концептуальні зв'язки, задані в онтології.

Реалізація цих вимог лежить у площині обов'язкового дотримання виконання загальної умови несуперечливості, що висувається до систем автоматичного опрацювання будь-якого виду інформації щодо форми структурованого представлення усіх типів даних, з якими працює комп'ютерна програма виконання операцій зіставлення.

У нашому випадку ключ до вирішення поставленого завдання (встановлення відповідності між одиницями СемП і ПО) знаходимо у створеній академіком В. А. Широковим загальній лексикографічній теорії знакових семантичних систем, функціональним призначенням яких є відображення об'єктів навколишнього світу, їхніх властивостей та зв'язків, що існують між ними [Широков 1998].

Семіотична сутність природних мов, за твердженням автора, дає право вважати їхні лексикографічні підсистеми частковим випадком зазначених систем.

Об'єднувальним моментом усіх лексикографічних систем є наявність у кожній з них так званого лексикографічного ефекту, який визначається як «відображення певного аспекту знань» про об'єкти навколишньої дійсності.

Як показали результати численних досліджень з лексичної семантики, зокрема й тих, що проводяться в Українському мовно-інформаційному фонді (УМІФі) під керівництвом В. А. Широкова в межах створення Інтегрованої лексикографічної системи української мови [Интегрированная лексикографическая...; Інтегрована...; Широков, Олександрук], найбільший вияв цього ефекту властивий великим тлумачним словникам (ТС), які на цей час є найрепрезентативнішим джерелом семантичної інформації, що забезпечує опис лексичного значення слова за багатьма параметрами із залученням прагматичного, конотативного, синтагматичного і лінгвістичного контекстів. Проведений змістовний аналіз структури тлумачних статей ТС виявив відображення в них тих самих закономірностей в організації передання значення слова, що й в інформаційних довідниках енциклопедичного типу, термінологічних тлумачних словниках, у тезаурусах, предметних онтологіях. [Дієслово...; Морковкин; Пшеничная; Сніжко]. Стосується це насамперед дотримання в тлумачному словнику принципу описування значення слів через встановлення зв'язків між референтами, позначуваними цими словами. Аналогами строго визначеного в спеціалізованих словниках логічного відношення «рід — вид» у тлумачному словнику виступають гіперо-гіпонімічні відношення. Усе це дає право говорити, по-перше, про доцільність проведення семантичних досліджень мовних одиниць (зокрема й пов'язаних із завданням експлікування поняттєвого змісту в лексичному значенні слова) опосередковано через аналіз тлумачень їхніх лексичних значень, заданих у

ТС¹, а по-друге, про коректність використання в ролі формального апарату для проведення такого аналізу формалізму, побудованого в межах теорії семантичних станів одиниць лексикографічних систем природних мов [Грама-тичні системи...; Широков 2005].

Теорія створена на тих самих засадах лексичної семантики, що й компонентний семантичний аналіз, і зорієнтована на використання комп'ютерних технологій у семантичних дослідженнях.

Згідно з цією теорією, окреме значення слова розглядається як його конкретний лінгвістичний стан, детермінований певним семантичним контекстом, представленим множиною ознак лексичної та граматичної семантики. Це положення є підставою для визнання одиницею семантичного аналізу багатозначних слів їхніх лексико-семантичних варіантів (ЛСВ). Однозначні слова вважаються як такі, що складаються з одного ЛСВ.

Формально відповідність між мовною одиницею та її семантичним станом описується як $g: S \rightarrow g(S)$, де S — лексико-семантичний варіант, $g(S)$ — семантичний стан ЛСВ, який має своїми детермінантами елементи засобів матеріального вираження семантики; g — відповідність між S та $g(S)$.

На множині семантичних станів Y задається оператор F , який інтерпретується як оператор сукупності значень певних семантичних категорій. Дія оператора F становить сумарну дію операторів M , T , X_i .

Оператор M вводить в описування значення слова компонент граматичної семантики у вигляді ознаки «лексико-граматичний клас лексеми».

За допомогою оператора T визначається семантична тема, яка вказує на поняттєву (сміслову) сферу, в якій відображаються певні типи реалій фізичного і ментального світу, до яких належать референти ЛСВ.

Оператори X_i репрезентують диференційні семантичні компоненти відповідної семантичної теми, її смислові характеристики.

Дія оператора F як певного інтелектуального механізму описується формулою $Fg_i(S) = \{M(L), T(S), X_i\}$ (1).

$T(S)$ встановлюються в межах значення оператора M . Наприклад, для M у значенні «прикметник» T може приймати значення: T_1 — квалітативна ознака, T_2 — квантитативна ознака, T_3 — реляційна.

Якщо T_i конкретизують загальне лексико-граматичне категоріальне значення слів однієї частини мови, окреслюючи в ньому певні поняттєві сфери, що співвідносяться з множиною взаємопов'язаних об'єктів позамовної дійсності, то оператори X_i уточнюють, про які саме конкретні характеристики позначуваних референтів ідеться при віднесенні їхніх значень до однієї семантичної теми.

Наприклад, видові ознаки семантичної теми T_1 «квалітативна ознака», визначеної в межах оператора M «прикметник», представлено множиною операторів $\{X_i\}$, які описують властивості, притаманні самим означуваним об'єктам безвідносно до їхніх зв'язків з іншими об'єктами: X_1 — носій ознаки, X_2 — можливість мати різний ступінь інтенсивності вияву, X_3 — ступінь вияву ознаки, X_4 — темпоральна характеристика, X_5 — розмір, X_6 — зовнішність/вигляд; властивості, що сприймаються органами відчуття: X_7 — на дотик (м'який, твердий), X_8 — на смак (гіркий, солодкий), X_9 — зір (колір);

¹ Ідеться про електронні версії ТС, побудовані з урахуванням спільної вимоги до всіх електронних словників щодо експліцитного й уніфікованого представлення в них інформації.

X_{10} — моральні якості, X_{11} — інтелектуальні якості, X_{12} — фізичний стан, X_{13} — психологічний стан, X_{14} — емоційний стан, X_{15} — світосприйняття, X_{16} — темперамент.

Зв'язки між операторами M_i (граматичне значення лексеми, його лексико-граматичний клас) і $\{T_i\}$ (семантична тема), між T_i і $\{X_i\}$ (сміслові ознаки теми) є відображенням семантичних гіперо-гіпонімічних відношень, встановлюваних на рівні денотативних значень слова. Це дає право говорити, що формула семантичного стану цілком відповідає положенням сучасної теорії лексичної семантики про компонентний склад лексичного значення слова і про ієрархічний характер відношень між визначеними компонентами. У формулі семантичного стану ці компоненти представлені конкретними значеннями операторів M_i , T_i , X_i . Задану у формулі ієрархію зв'язків між операторами можемо зобразити схематично (див. рис. 1)

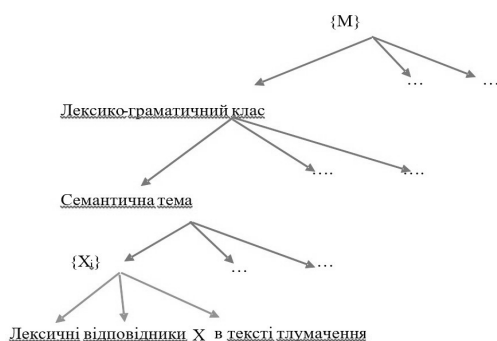


Рис. 1. Ієрархія зв'язків між операторами

Опис лексичного значення однозначного слова або ЛСВ багатозначного слова за допомогою формули семантичного стану виконується в стандартизований процедурний спосіб, який полягає в тому, щоб виявити в словниковому тлумаченні лексемні відповідники значенням операторів M , T , X у формулі. Уніфіковане таким чином представлення лексичних значень слів відкриває шлях до здійснення над ними в автоматизованому режимі цілого ряду логічних операцій класифікаційного плану. Семантична класифікація може при цьому проводитися, по-перше, за кожним окремим параметром або за сукупністю параметрів, охоплених формулою семантичного стану, визначеною в межах конкретного значення оператора M . Наприклад, за параметром «носії ознаки» з формули семантичного стану якісних прикметників можуть бути сформовані групи прикметників на позначення властивостей: людей, представників тваринного світу, фізичних тіл, явищ природи тощо; за кон'юнкцією параметрів {«носії ознаки» $\wedge$ «позитивний характер ознаки» $\wedge$ «ступінь прояву ознаки» $\wedge$ «моральні якості»} можна виділити семантичну групу прикметників на позначення характеристик високоморальних людей. Чим більше залучено X_i у множині однакових ознак, за якими здійснюється класифікація, тим тісніший зв'язок між елементами утворених семантичних груп.

У результаті зіставлення структурних елементів предметних онтологій і структурної формули семантичного стану (1), за допомогою якої експлікуються семантичні складники лексичного значення слова, встановлюється така позиційно-значеннєва відповідність між елементами цих структур:

семантичним темам $\{T_i\}$, у яких експлікований поняттєвий зміст значення слова, в ПО відповідають концепти; смисловим характеристикам теми $\{X_i\}$ з формули (1) — атрибуту концептів у ПО. Задана у формулі семантичного стану ієрархія зв'язків між операторами T (родова ознака) і X (видова ознака) є відображенням концептуальних відношень, які в ПО встановлюються у результаті аналітичного опрацювання знань про іманентні властивості предметів (об'єктів).

Встановлені відповідності свідчать про ефективність застосування формалізму теорії семантичних станів не тільки для представлення номінативної інформації, що міститься в лексичному значенні мовних одиниць, та процедурного здійснення операцій порівняльного аналізу різного роду, але й під час вирішення завдань, пов'язаних з аналізом мовної інформації на рівні глибоких смислових структур концептуального плану.

Для здійснення автоматизації обох видів аналізу (семантичного і семантико-когнітивного), які передбачають залучення інформації про семантичні компоненти лексичного значення мовних одиниць, встановлювані за допомогою формули семантичного стану (ФСС), необхідно, щоб значення операторів у ФСС і їхні лексемні відповідники з тексту², що аналізується, були представлені в кодах семантичних класів, які об'єднують у собі слова однієї поняттєвої категорії з однаковим лексичним значенням.

У термінах теорії семантичного стану визначення семантичного класу (СК) може бути інтерпретоване як множина слів або/ї їхніх ЛСВ, які за формулою належать до однієї лексико-граматичної категорії M_i , уживаються на позначення того самого поняття T_i і об'єднані за спільною семантичною характеристикою X_i . Тоді, згідно з цим визначенням, ознаці «інтелектуальні якості» — X_{11} у ФСС прикметників — відповідатимуть семантичні класи: «розумові здібності», «творча обдарованість», «винахідливість». Ідентифікаторами першого з них у тексті тлумачень є слова: *розум, інтелект, кмітливість, кмітливий, тямущий, розсудливість, розсудливий, глузд, освічений, освіченість, учений, вчений, тямкий, мислячий, знання, тямущість*; другого — *майстерність, здібність, здібний, обдарований, обдарованість, творчий, талант, геній, геніальний, винятковий*; третього — *винахідливий, винайдений, придуманий, вигадливий, вигадливість, вимислений, вимислити, кмітливий, кмітливість*.

У результаті заміни в ФСС значень її операторів T і X , виражених словами, на коди семантичних класів, до яких належать ці слова, отримуємо уніфіковане експліковане представлення семантичних компонентів лексичного значення слова у формі ланцюжка числових кодів семантичних класів і тим самим робимо можливим залучення комп'ютера до роботи з семантико-когнітивною інформацією.

Нижче наведено приклад цифрового запису формул семантичного стану прикметника *розумний* — ‘який має високий ступінь інтелекту, здатний логічно мислити’, прикметника *кмітливий* — ‘здатний добре і швидко міркувати’, прикметника *синій* — ‘який має забарвлення одного з основних кольорів спектра — середній між голубим і фіолетовим’, прикметника *блакитний* — ‘небесно-голубого кольору’:

² У цьому випадку тлумачна стаття в ТС, одиниці реєстрової частини словника, взагалі будь-яке висловлення, написане знаками природної мови, — це тексти.

розумний — {2; 210; 220, 230, 101, 1000, 1001, 1005, 1010}
 кмітливий — {2; 210; 220, 230, 101, 1000, 1001, 1005}
 синій — {2; 210; 220, 301, 3000, 3005, 3009}
 блакитний — {2; 210; 220, 301, 3000, 3020, 3025}.

У представленому цифровому записі значення кодів такі: 2 — прикметник (М: лексико-граматичний клас), 210 — ознака (Т — загальна концептуально-мовна категорія); значення оператора Х: 220 — квалітативна характеристика, 230 — позитивна характеристика, 101 — предмет: істота (носії ознаки), 1000 — інтелектуальна характеристика, 1001 — наявність ступеня прояву, 1005 — високий ступінь прояву, 301 — предмет неживої природи (носії ознаки), 3000 — забарвлення, 3005 — належність до основного спектра кольорів, 3009 — місце в спектрі: між блакитним і фіолетовим, 3020 — неосновний колір спектра (відтінок), 3025 — невеликий ступінь прояву основного кольору.

Семантичний класифікатор будується як рівнева структура, яка є відображенням відношень, що встановлюються між семантичними компонентами у ФСС (див. рис.2).

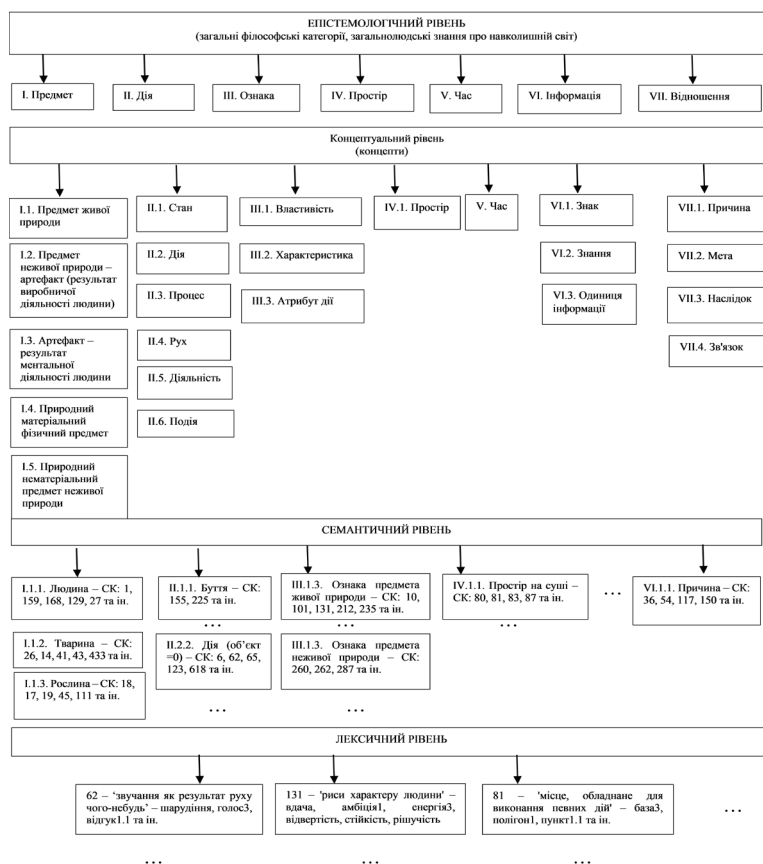


Рис. 2. Семантичний класифікатор (фрагмент)

Епістемологічний рівень відображає загальні філософські категорії, загальнолюдські знання про навколишній світ. Концептуальний рівень (концепти) забезпечує для кожної лексичної одиниці операційне визначення її

місця в поняттєвій системі картин світу. На семантичному рівні перебувають семантичні класи, об'єднані в групи за спільною семантичною темою, визначеною з формули семантичного стану лексичних репрезентантів відповідних класів:

I.1.1. Людина – СК: 1, 159, 168, 129, 27 та ін.

I.1.2. Тварина – СК: 26, 14, 41, 43, 433 та ін.

I.1.3. Рослина – СК: 18, 17, 19, 45, 111 та ін.

II.2.2. Дія (об'єкт = 0) – СК: 6, 65, 62, 618, 123 та ін.

...

III.1.3. Ознака предмета живої природи – СК: 10, 101, 131, 212, 235 та ін.

...

IV.1.1. Простір на суші – СК: 80, 81, 83, 87 та ін.

...

VI.1.1. Причина – СК: 36, 54, 117, 150 та ін.

Лексичний рівень (лексемне наповнення семантичних класів) містить інформацію про числовий код СК, значення X_i , за яким слова, взяті зі «Словника української мови» в 20 томах [Словник], об'єднуються в семантичну групу СК, лексемні репрезентанти (ЛСВ) семантичного класу, наприклад:

81 — 'місце, обладнане для виконання певних дій': *база3, полігон1, цвинтар1, пункт1.1* та ін.;

62 — 'звучання як результат руху чого-небудь': *шарудіння, скрип, голос3, відгук1.1* та ін.;

131 — 'риси характеру людини': *вдача, амбіція1, енергія3, відвертість, стійкість, рішучість* та ін.

Запропонований у ролі формального апарату семантичного аналізу формалізм теорії семантичного стану в поєднанні з семантичним класифікатором, за допомогою якого здійснюється уніфікована інтерпретація операторів формули семантичного стану і семантична індексація текстових слів, знайшов своє практичне застосування в дослідженнях із семантики, які проводяться в УМІФі в межах розв'язання глобальної проблеми автоматизованої переробки мовної інформації засобами інтегрованих лексикографічних систем. Прикладами таких застосувань можуть служити побудова за допомогою цього апарату операційного критерію встановлення семантичної подібності між ЛСВ різних слів при формуванні та ранжуванні синсетів «Електронного словника синонімів української мови» [Грязнухіна, Устимець, Широков], а також здійснення за допомогою формалізму аналізу семантики відносних прикметників, заданих у ТС імпліцитними формулами тлумачення «Прикм. до...», «Стос. до...», і побудова експліцитного опису їхніх лексичних значень шляхом включення у відповідні цим прикметникам формули семантичного стану характеристик семантичної валентності [Грязнухіна, Любченко, Олександрук; Широков, Олександрук].

Результати дослідження дають змогу зробити висновок про доцільність застосування мовно-інформаційного інструментарію для вирішення таких важливих завдань комп'ютерної лінгвістики, як визначення формальних критеріїв розмежування лексичних омонімів і ЛСВ багатозначного слова, коректне встановлення структурних протиставлень «значення — відтінки», «лексема — лексико-семантичні варіанти» в тлумаченні значення останнього, завдання уніфікації опису лексичного значення слів одного семантичного поля. Розроблений мовно-інформаційний інструментарій також можна

застосовувати під час розв'язання проблем, пов'язаних з автоматизацією екстрагування семантичної інформації з текстів та подальшого її формалізованого аналізу в плані зіставлення значень текстових одиниць з метою встановлення факту їхньої семантичної подібності чи відмінності.

ЛІТЕРАТУРА

- Апресян Ю. Д. Избранные труды. 2-е изд., испр. и доп. Москва, 1995. Т. 1. Лексическая семантика. 472 с.
- Грамматичні системи: феноменологічний підхід / В. А. Широков, Т. П. Любченко, І. В., Шевченко, К. В. Широков. Київ : Наук. думка, 2018. 310 с.
- Грязнухіна Т. О., Любченко Т. П., Олександрук І. В. Феноменологічний підхід до вивчення семантики відносних прикметників та способу представлення її в електронному тлумачному словнику. *Актуальні проблеми філології та перекладознавства* : зб. наук. пр. Хмельницький, 2019. Вип. 16. С. 39–44.
- Грязнухіна Т. О., Устимець О. В., Широков В. А. Операційне визначення критерію семантичної подібності синонімів. *Прикладна лінгвістика та лінгвістичні технології: MegaLing-2008* : зб. наук. пр. Київ : Довіра, 2009. С. 42–57.
- Дарчук Н. П. Комп'ютерне анотування українського тексту: результати і перспективи. Київ : Освіта України, 2013а. 543 с.
- Дарчук Н. П. Напрямок формалізації семантики. *Мовні і концептуальні картини світу*. Київ, 2013б. Вип. 46(1). С. 385–396.
- Дієслово в лексикографічній системі / О. Г. Рабулець, Н. М. Сухарина, В. А. Широков, К. М. Якименко. Київ : Довіра, 2004. 259 с.
- Интегрированная лексикографическая система как инструмент автоматизированного формирования новых электронных словарей и как база лингвистических исследований в области лексической семантики / Т. А. Грязнухина, Н. М. Заика, Т. П. Любченко, В. А. Широков. *Актуальные проблемы современной прикладной лингвистики* : сб. науч. статей. Минск : МГЛУ, 2017. С. 221–225.
- Інтегрована лексикографічна система «Словники України», версія 1.03 / В. А. Широков, О. Г. Рабулець, І. В. Шевченко, О. М. Костишин, К. М. Якименко. Електронний оптичний диск (CD-ROM).
- Крушевский Н. В. Очерк науки о языке. Казань, 1883. Т. 19. 148 с.
- Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы : учеб. пособие. Москва : Академия, 2006. 303 с.
- Лінгвістично-інформаційні студії : пр. Укр. мов.-інформ. фонду НАН України : в 5 т. / за заг. ред. В. А. Широкова. Київ : УМІФ, 2018. Т. 1 : Наукова парадигма та основні мовно-інформаційні структури. 271 с.
- Любченко Т. П. Лексикографічні системи граматичного типу та їх застосування в засобах автоматизованого опрацювання мови : автореф. дис. ... канд. техн. наук. Київ, 2011. 19 с.
- Морковкин В. В. Идеографические словари. Москва : Изд-во : МГУ. 1970. 72 с.
- Морфологический анализ научного текста на ЭВМ / В. И. Перебейнос, Т. А. Грязнухина, Н. П. Дарчук и др.. Киев : Наук. думка, 1989. 262 с.
- Пиотровский Р. Текст, машина, человек. Ленинград : Наука, 1975. 327 с.
- Пшеничная Л. Э. Тезаурус в документальной ИПС. Київ : Наук. думка, 1977. 121 с.
- Рубашкин В. Онтология как инструмент концептуальной интерпретации лексики. *Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011* : зб. наук. пр. Київ : УМІФ, 2012. С. 318–330.
- Синтаксический анализ научного текста на ЭВМ / Т. А. Грязнухина, Н. П. Дарчук, В. И. Критская и др. Киев : Наук. думка, 1999. 272 с.
- Словник української мови online. Т. 1–13 (А–ПОКІРНО). URL: <https://sum20ua.com/>

- Entry/index?wordid=17316&page=580 (дата звернення: 21.02.2023).
- Сніжко Н. В. Идеографичний тезаурус як модель лексико-семантичної системи (за наслідками автоматизованого аналізу українських іменників). *Мовознавство*. 1995. № 6. С. 28–35.
- Уемов А. И. Системный подход и общая теория систем. Москва : Мысль, 1978. 272 с.
- Шевченко І. В. Моделі та алгоритмічно-програмне забезпечення лексикографічних систем : автореф. дис. ... канд. техн. наук. Київ, 2000. 19 с.
- Широков В. А. Інформаційна теорія лексикографічних систем. Київ : Довіра, 1998. 331 с.
- Широков В. А., Олександрук І. В. Віртуальна лексикографічна лабораторія «Словник української мови» як інструмент семантичних досліджень. *Мовознавство*, 2019. № 5. С. 3–20.
- Широков В. А. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії. *Мовознавство*. 2005. №3–4. С. 47–62.
- An Outline of Linguistic Science [«Очерк науки о языке» в переводе на английский язык] / M. Kruszewski Writings in General Linguistics; ed. with an Introduction by K. Koerner. Amsterdam; Philadelphia : John Benjamins, 1995. 188 p.
- Nirenburg S., Raskin V. Ontological Semantics. Cambridge, MA : MIT Press, 2004. 420 p.
- Staab S., Studer R. Handbook on Ontologies. Berlin; Heidelberg : Springer-Verlag, 2004. 500 p.

Дата надходження до редакції: 05.04.2023

Дата надходження після доопрацювання: 22.06.2023

Дата затвердження редакцією: 26.06.2023

REFERENCES

- An Outline of Linguistic Science [«Ocherk nauki o jazyke» v perevode na anglijskij jazyk]. (1995). M. Kruszewski Writings in General Linguistics. K. Koerner (Ed., Intro.). Amsterdam; Philadelphia: John Benjamins.
- Apresyan Ju. D. (1995). Izbrannye trudy. 2nd ed. Vol. 1. Leksicheskaya semantika. Moscow. [In Russian].
- Gryaznukhina T. O., Lyubchenko T. P., Oleksandruk I. V. (2019). Fenomenologichni pidkhid do vyvchennia semantyky vidnosnykh prykmetnykiv ta sposobu predstavlennia yiyi v elektronnomu tlumachnomu slovnyku. *Aktualni problemy filolohii ta perekladoznavstva. Zbirnyk naukovykh prats*. Khmelnytskyi, (16), 39–44. [In Ukrainian].
- Gryaznukhina T. O., Ustymets O. V., Shyrokov V. A. (2009). Operatsiine vyznachennia kryteriiu semantychnoi podobnosti synonimiv. *Prykladna linhvystyka ta linhvystychni tekhnolohii : MegaLing-2008. Zbirnyk naukovykh prats*. Kyiv : Dovira, 42–57. [In Ukrainian].
- Hramatychni systemy: fenomenologichni pidkhid. (2018). V. A. Shyrokov, T. P. Lyubchenko, I. V. Shevchenko, K. V. Shyrokov. Kyiv : Naukova dumka. [In Ukrainian].
- Darchuk N. P. (2013). Komp'uterne anotuvannia ukraïnskoho tekstu: rezultaty i perspektyvy. Kyiv : Osvita Ukrainy. [In Ukrainian].
- Darchuk N. P. (2013). Napriamky formalizatsii semantyky. *Movni i kontseptualni kartyny svitu*. Kyiv, 46(1), 385–396. [In Ukrainian].
- Diieslovo v leksykografichnii systemi. (2004). O. H. Rabulets, N. M. Sukharyna V. A. Shyrokov, K. M. Yakymenko. Kyiv : Dovira. [In Ukrainian].
- Integrirovannaya leksikograficheskaya sistema kak instrument avtomatizirovannogo formirovaniya novykh elektronnykh slovarej i kak baza lingvisticheskikh issledovanij v oblasti leksicheskoy semantiki. (2017). T. A. Gryaznukhina, N. M. Zaika, T. P. Lyubch

- enko, V. A. Shirokov. *Aktual'nye problemy sovremennoj prikladnoj lingvistiki: sbornik nauchnykh statej*. Minsk : MGLU, 221–225. [In Russian].
- Intehrovana leksykohrafichna systema «Slovnyky Ukrainy», versia 1.03. V. A. Shyrokov, O. H. Rabulets, I. V. Shevchenko, O. M. Kostyshyn, K. M. Yakymenko. CD-ROM. [In Ukrainian].
- Krushevskij N. V. (1883). *Ocherk nauki o yazyke*. Kazan, (19). [In Russian].
- Leont'eva N. N. (2006). *Avtomaticheskoe ponimanie tekstov: sistemy, modeli, resursy: uchebnoe posobie*. Moscow : Akademiya. [In Russian].
- Linhvistychno-informatsiini studii: pratsi Ukrainskoho movno-informatsiinoho fondu NAN Ukrainy : in 5 vols. (2018). Vol. 1. *Naukova paradyhma ta osnovni movno-informatsiini struktury*. V. A. Shyrokov (Chief Ed.). Kyiv. [In Ukrainian].
- Lyubchenko T. P. (2011). *Leksykohrafichni systemy hramatychnoho typu ta yikh zastosuvannia v zasobakh avtomatyzovanoho opratsiuvannia movy*. Abstract of Cand. of Techn. Sc. Diss. Kyiv. [In Ukrainian].
- Nirenburg S., Raskin V. (2004). *Ontological Semantics*. Cambridge, MA: MIT Press.
- Morkovkin V. V. (1970). *Ideograficheskie slovari*. Moscow : Izdatel'stvo MGU. [In Russian].
- Morfologicheskij analiz nauchnogo teksta na EVM. (1989). V. I. Perebeynos, T. A. Gryaznuhina, N. P. Darchuk et al. Kyiv : Naukova dumka. [In Russian].
- Piotrovskij P. (1975). *Tekst, mashina, chelovek*. Leningrad : Nauka. [In Russian].
- Pshenichnaya L. E. (1977). *Tezaurus v dokumental'noj IPS*. Kyiv : Naukova dumka. [In Russian].
- Rubashkin V. *Ontologiya kak instrument kontseptual'noj interpretatsii leksiki*. (2012). *Prykladna linhvistyka ta linhvistychni tekhnolohii: MegaLing-2011 : Zbirnyk naukovykh prats*. Kyiv, 318–330. [In Russian].
- Sintaksicheskij analiz nauchnogo teksta na EVM. (1999). T. A. Gryaznukhina, N. P. Darchuk, V. I. Kritskaya et al. Kyiv : Naukova dumka. [In Russian].
- Shevchenko I. V. (2000). *Modeli ta alhorytmichno-prohramne zabezpechennia leksykohrafichnykh system*. Abstract of Cand. of Techn. Sc. Diss. Kyiv. [In Ukrainian].
- Shyrokov V. A. (1998). *Informatsiina teoriia leksykohrafichnykh system: monograph*. Kyiv : Dovira. [In Ukrainian].
- Shyrokov V. A., Oleksandrak I. V. (2019). *Virtualna leksykohrafichna laboratoriiia «Slovnyk ukrainskoi movy» yak instrument semantychnykh doslidzhen*. *Movoznavstvo*, (5), 3–20. [In Ukrainian].
- Shyrokov V. A. (2005). *Semantychni stany movnykh odynyt ta yikh zastosuvannia v kohnityvnii leksykohrafii*. *Movoznavstvo*, (3–4), 47–62. [In Ukrainian].
- Slovnyk ukrainskoi movy online. Vol. 1–13 (A–POKÍRNO). URL: <https://services.ulif.org.ua/expl/Entry/index?wordid=1&page=0> (last accessed: 21.02.2023). [In Ukrainian].
- Snizhko N. V. (1995). *Ideografichniy tezaurus yak model leksyko-semantychnoi systemy (za naslidkamy avtomatyzovanoho analizu ukrainskykh imennykiv)*. *Movoznavstvo*, (6), 28–35. [In Ukrainian].
- Staab S., Studer R. (Eds). (2004). *Handbook on Ontologies*. Berlin; Heidelberg : Springer-Verlag.
- Uemov A. I. (1978). *Sistemnyj podhod i obshchaya teoriya sistem*. Moscow : Mysl'. [In Russian].

Received: 05.04.2023

Received in revised form: 22.06.2023

Accepted: 26.06.2023

Tetyana GRYAZNUKHINA

Ukrainian Lingua-Information Foundation of the National Academy of Sciences of Ukraine

3 Holosiivskyi Ave., Kyiv, 03039, Ukraine

E-mail: umiftatyana@gmail.com

<https://orcid.org/0000-0002-8055-442X>

Tetyana LYUBCHENKO

Ukrainian Lingua-Information Foundation of the National Academy of Sciences of Ukraine

3 Holosiivskyi Ave., Kyiv, 03039, Ukraine

E-mail: t.lyubch@gmail.com

<https://orcid.org/0000-0002-7390-0427>

Iryna OLEKSANDRUK

O. O. Potebnia Institute of Linguistics of the National Academy of Sciences of Ukraine

4 Hrushevskoho Str., Kyiv, 01001, Ukraine

E-mail: ira_oleksandruk@ukr.net

<https://orcid.org/0000-0003-2701-1714>

DEVELOPMENT OF LANGUAGE AND INFORMATION TOOLS FOR AUTOMATED EXTRACTION OF SEMANTIC AND COGNITIVE INFORMATION FROM THE TEXT

The article deals with the role of semantic research for the creation of intelligent systems for automated processing of language information. An approach to the development of linguistic and informational tools for automated extraction of semantic and cognitive information from the text is outlined. The issue of increasing the efficiency of high-quality and fast processing of digital information, the creation of highly intelligent, fast-acting information systems was raised.

The theoretical basis of the research is the lexicographic theory of symbolic semantic systems developed by academician V. A. Shyrokov and the theory of semantic states of natural language units built within it, which is based on the principles of lexical semantics and is oriented towards the use of computer technologies in semantic research.

An integrated lexicographic system created in the Ukrainian Language and Information Fund of the National Academy of Sciences of Ukraine was selected as the source base of the study, the core component of which is the computer version of the «Dictionary of the Ukrainian Language» in 20 volumes, which is the most representative source of semantic information that provides a description of the lexical meaning of a word according to many parameters with the involvement of pragmatic, connotative, syntagmatic and linguistic contexts.

In the study, the formalism of the theory of semantic states is used to build a linguistic and informational toolkit for automatic recognition of semantic and cognitive information in texts written in natural languages.

The proposed semantic classifier provides an explicit presentation of the information extracted from the text, which enables its further software processing thanks to the unified interpretation of the operators of the semantic state formula and the results of the semantic indexing of the text.

The expediency of using the developed language and information toolkit for solving such important tasks of computer linguistics as determining formal criteria for distinguishing lexical homonyms and LSV of a polysemous word, correct establishment of structural contrasts «meaning — shades», «lexeme — lexical-semantic variants» in interpreting the meaning of the latter, the task of unifying the description of the lexical meaning of words of one semantic field.

Keywords: automatic information processing, theory of symbolic lexicographic systems, integrated lexicographic system, semantic state formula, semantic and cognitive information.