

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
УКРАЇНСЬКИЙ МОВНО-ІНФОРМАЦІЙНИЙ ФОНД

В. А. ШИРОКОВ

ЕЛЕМЕНТИ ЛЕКСИКОГРАФІЇ

КИЇВ
ВИДАВНИЦТВО "ДОВІРА"
2005

У монографії викладаються результати багаторічної діяльності Українського мовно-інформаційного фонду Національної академії наук України зі створення національної словникової бази. Подано нову концепцію та комп'ютерну технологію лексикографічної діяльності. На основі аналізу системотвірних відношень та інваріантів мовної субстанції побудовано формальні засади теорії лексикографічних систем, яка надає концептуальні засоби для створення і застосування словників та словникових комплексів у системах комп'ютерного опрацювання природномовної інформації. Розглянуто численні застосування теорії лексикографічних систем на реальних прикладах створення об'єктів національної словникової бази, серед яких вирізняються електронні граматичні словники, інтегрована лексикографічна система "Словник України", інструментальна система підтримки фундаментальної тлумачної лексикографії та ряд інших. На основі розвинутої автором теорії семантичних станів обґрунтовано й розроблено концептуальні і системотехнічні засади побудови багатомовних словникових систем та віртуальних систем професійної взаємодії в лінгвістиці.

Відповідальні редактори:
Т. О. Грязнухіна, О. Г. Рабулець

Рецензенти:
академік НАН України *В. М. Русанівський*,
член-кореспондент НАН України *О. В. Палагін*

*Рекомендовано до друку
вченою радою Українського мовно-інформаційного фонду
НАН України*

ВСТУП

Цивілізаційне значення процесів функціонування знання в постіндустріальному суспільстві та роль мови в цих процесах набувають останнім часом таких вимірів, що виводять мовознавство з кола суто гуманітарних наукових дисциплін і надають йому якостей дисципліни технологічної, спонукаючи до висновку, що в постіндустріальних умовах природна людська мова — мабуть уперше в історії людської цивілізації — набуває технологічного статусу¹, від якого починає безпосередньо залежати ефективність функціонування виробництва, а зрештою — й інших суспільних інститутів. При конструюванні сучасних інформаційно-комунікаційних технологій постала потреба у врахуванні та застосуванні фундаментальних властивостей мовної субстанції, маючи на меті створення мовно-інформаційних артефактів, налаштованих на інтелектуальне опрацювання мови і потрібних для функціонування високоєфективних технологій оперування знаннями. Звернімося до аналізу особливостей функціонування мови в інформаційному суспільстві та суспільстві знань.

Еволюція постіндустріального суспільства привела до того, що останнім часом замість поняття “інформаційне суспільство” все частіше використовуються “суспільство знань” або “суспільство, орієнтоване на знання”; все активніше застосовуються терміни типу “економіка, заснована на знаннях” та “економіка знань”. Різні країни та їх об'єднання обирають парадигму знання за основу своєї національної, а подекуди вже й інтернаціональної стратегії — згадаймо у цьому зв'язку хоча б проєкт побудови так званого “European Knowledge Society”². Отже, принаймні передові країни світу вже впевнено перебувають на етапі переходу до другої фази

¹ Широков В.А. Гуманітарна традиція і технологічний статус мови. Мовознавство, 2001, №3.

² European Knowledge Society www.eurofound.eu.int/areas/industrialchange/knowledgesociety/
<http://www.meaningprocessing.com/personalPages/tuomi/articles/TheFutureOfTheEuropeanKnowledgeSociety/etc>.

постіндустріального суспільства, власне — до суспільства знань. Не минули ці віяння і Україну, яка на державно-урядовому рівні останнім часом також неодноразово декларувала відданість знаньорієнтованому шляхові розвитку³.

Зрозуміло, що кожному суспільно-виробничому укладу відповідає характерний для нього комплекс технологічних інструментів. Зокрема, для інформаційного суспільства це були інформаційно-комунікаційні технології. Для суспільства знань це, очевидно, мусять стати технології оперування знаннями, причому на всьому їх життєвому циклі: від створення знань до їх впровадження в економічно-виробничу систему — тобто саме туди, де вони дещо таємничим способом перетворюються на конкретну продукцію. Іншими словами, зараз маємо об'єктивну потребу поставити на технологічну основу процеси опрацювання знань, тобто — створити ефективні технології опрацювання знань.

Для розробки таких технологічних схем та засобів необхідно, насамперед, мати визначення поняття знання, яким можна було б оперувати не у філософських цілях, а в робочому, прагматичному режимі. Відтак, потрібні дефініції знання, спроможні відіграти роль об'єкта технологічного моделювання та конструювання. У зв'язку з викладеним зробимо декілька зауважень, так би мовити, фізичного та метафізичного характеру, які допоможуть увійти до кола понять, дотичних до предмета, що розглядається.

По-перше, інтуїтивно зрозуміло, що поняття знання кореспондується з поняттям інформації. Але по суті вони є цілком різними. Інформація є об'єктивною характеристикою об'єктивних процесів. На нашу думку, існує глибока аналогія між визначеннями понять інформації, з одного боку, та енергії, яка також є певною об'єктивною характеристикою матерії — з іншого. Цю аналогію було простежено у низці праць, у тому числі і в наших книгах⁴. За сучасними науковими уявленнями інформація являє собою об'єктивну властивість будь-яких об'єктів, систем, відношень, процесів і характеризує такі їх якості як структурованість, неоднорідність, складність. Апофеозом цього підходу є алгоритмічна теорія інформації А.М.Колмогорова, яка, власне, ґрунтується на уявленні та математичній формалізації поняття складності. Зазначена

³ Економіка знань: виклики глобалізації та Україна. Під заг.ред. А.С.Гальчинського, С.В.Львовичкіна, В.П.Семиноженка. — К.: Національний інститут стратегічних досліджень. 2004. 261 с.; Україна на шляху до суспільства знань. — К.: Прайвесі Юкрейн. 2005. 69 с.

⁴ В.А.Широков. Інформаційна теорія лексикографічних систем. — К.: Довіра, 1998. 331 с. В.А.Широков. Феноменологія лексикографічних систем. — К.: Наукова думка, 2004. 326 с.

теорія подає й кількісну міру інформації — так звану алгоритмічну інформаційну міру Колмогорова. Оскільки, структурованість, неоднорідність та складність тією чи іншою мірою притаманні будь-яким об'єктам, системам, відношенням чи процесам, то очевидно, що і інформація є однією з їхніх універсальних об'єктивних характеристик. Питання полягає лише у встановленні формалізованих моделей інформації, адекватних тим чи іншим класам процесів, а також у розробці відповідних методів та засобів її кількісного оцінювання та вимірювання. З оглядом основних понять та ідей інформаційної теорії можна ознайомитися у роботах^{4, 5, 6} та посиланнях, що в них містяться.

Постійне підкачування нової інформації є необхідною умовою інтенсивного, або як говорять останнім часом, сталого розвитку економіки і суспільства взагалі.

Але чи є ця умова достатньою?

Відповідь на це питання є негативною.

Справді, далеко не будь-яку інформацію можна перетворити на корисний ресурс. У прикладах, розглянутих у працях⁴, такі перетворення в реальних виробничо-технологічних ситуаціях виконують цілі комплекси організацій та інституцій — хочеться, слідуючи за Селфриджем⁷, вжити слово "пандемоніуми", які забезпечують продукування й багаторазові перетворення інформації та її транспортування, аж допоки вона не потрапить до соціотехнічної системи у потрібний момент та ще й у потрібній формі, адаптованій до сприйняття зазначеною системою.

Саме цей процес продукування та цілеспрямованого багаторазового перетворення і транспортування інформації надає їй таких якостей, котрі дозволяють нам кваліфікувати її як знання.

Отже, тепер можна дати таке робоче визначення поняття знання.

Знання — це інформація, форма якої є носієм трансформацій, котрим вона піддається в соціальній системі.

⁵ Szillard L. Über die Entropievermindung in einem Thermodynamischen System bei Eingriffen intelligenter Wesen. — Zs. f. Phys., 1929. — 53, 11-12. Heft. — P. 840-856. Стратонович Р.Л. Теория информации. — М.: Сов. радио, 1975. — 423 с., гл 12. Колмогоров А.Н. Три подхода к определению понятия "количество информации". // "Теория информации и теория алгоритмов". — М.: Наука, 1987. — С. 213-223. Шилейко А.В., Кочнев В.Ф., Хилушин Ф.Ф. Введение в информационную теорию систем. — М.: Радио и связь, 1985. — 278 с.

⁶ Волькенштейн М.В. Теория информации и эволюция. // "Кибернетика живого: Биология и информация". — М.: Наука, 1984. — С. 45-53. Волькенштейн М.В. Энтропия и информация. — М.: Наука, 1986. — 190 с.

⁷ Selfridge O.G. Pandemonium: a paradigm for learning. In: Mechanisation of thought processes. London. HMSO., 1959. — P. 511-531.

Підкреслимо, що це визначення не претендує на абсолют і є саме робочим, навіть технічним. Незважаючи на таку ланідарність з цієї дефініції випливають цілком конкретні наслідки.

По-перше, якщо інформація, як така, представляє певні об'єктивні властивості речей, то знання несе в собі потенціал суб'єктивного. Справді, перш ніж втрапити до виробничо-економічної системи, інформація повинна стати фактом свідомості — спочатку індивідуальної, а потім і колективної, принаймні групової. При цьому вона, зрозуміло, зазнає низку перетворень у відповідності зі специфікою функціонування свідомості. А ця специфіка є такою, і, відповідно, такою є конструкція людського інтелектуального апарату, що в ньому відбуваються численні перетворення та взаємодії між ментальними і мовними структурами.

Дослідження великих психологів і лінгвістів, серед яких згадаємо В.Бехтерева, Л.Виготського, А.Лурія, А.Леонтєва, Вільгельма фон Гумбольдта, О.Потебню, Р.Шенка та багатьох інших, переконливо підтвердили, що будь-який ментальний процес у людини має свою рефлексію в мовній сфері. Так що коректно говорити про єдиний мовнорозумовий процес. Підкреслимо також і ту важливу роль, яку відіграє спеціалізація мовної підсистеми як основного комунікативного чинника у людському суспільстві.

Отже, зі сказаного випливає ще одна (і досить очевидна) дефініція знання як інформації, вербалізованої та структурованої згідно із законами мовної системи. Адже в мовній формі сконцентровано такі важливі аспекти людського буття як культурний код, наукова та мовна картини світу, мовна свідомість та підсвідомість (колективна та індивідуальна) тощо.

Таким чином, наукова або контекстно-предметна картина світу, котра зазвичай є первинною в інформації, перетворюючись на суспільно значуще знання, активно взаємодіє з мовною системою і набирається від неї властивостей, притаманних цій системі. Взагалі, мовна та контекстно-предметна складові становлять два доволі відмінні аспекти функціонування знання. Причому саме мовна форма превалює на значному відрізку функціонування знання, аж до акту його безпосереднього споживання виробничою системою.

Відзначимо, що мовна форма, яку набуває знання, сама по собі є доволі зручною та гнучкою.

По-перше, вона є досить універсальною — маємо переконання, що майже будь-яка інформація може бути вербалізована.

По-друге, кожна людина досить кваліфіковано оперує цією формою, навчаючись неї із самого дитинства.

По-третє, мовна форма є лінійною, внаслідок чого вона легко піддається різним інформаційним операціям — кодуванню, перетворенню, зберіганню, транспортуванню і таке інше.

Але оберненою стороною цієї зручності та гнучкості виявляється те, що у природномовній формі інформації явно маніфестовано саме елементи мовної системи, тим часом як онтологічну, або семантичну складові представлено лише імпліцитно. Ще одне, важливе з нашої точки зору зауваження, впливає з природи знання як інформації, форма якої є носієм трансформацій, котрим вона піддається в соціальній системі. Адже ця форма (*природномовна форма*), будучи носієм соціальних трансформацій, водночас несе в собі і зміст суспільних процесів, причому, як історично тягла субстанція, вона імпліцитно містить в собі й суспільно-історичну складову — сліди численних, часто непростежуваних контактів з іншими мовами, численні когнітивні, психологічні та культурно-реліктові структури. З викладеного логічно випливає, що “витягування” семантичної інформації з вербалізованого знання, тобто природномовного тексту неминуче веде до застосування принципів мовної системи, так би мовити, в “оберненому” порядку. А саме, якщо символічно зобразити процес вербалізації інформації, як:

$$MI = T, \quad (B.1)$$

де M — “оператор” мовної системи, I — інформація, T — природномовний текст, то видобування семантичної інформації виглядатиме як:

$$M^{-1} T = I^{\text{сем}}, \quad (B.2)$$

де M^{-1} — оператор, обернений до M .

Останній процес ми й інтерпретуватимемо як екстракцію знань із тексту.

Зрозуміло, що написати формули (B.1)—(B.2) набагато легше ніж організувати відповідні процеси. Останнім часом спостерігається значний потік наукових публікацій на дану тему. Чимало з них присвячено методам автоматизованої побудови певних моделей знань за природномовним текстом. Стандартними моделями знань, як відомо, є фреймові, продукційні або логічні (дехто з дослідників кваліфікує їх як різні), а також модель семантичних мереж. Тим часом, великих успіхів на цьому шляху поки що немає. Причини цього, на нашу думку, є такими.

Насамперед, слід відзначити, що всі методики екстракції знань є дуже мовнозалежними, тобто вони залежать від конструкції системи конкретної національної мови, якою представлено вхідну інформацію. Тому для переходу до інтелектуального аналізу тексту для

кожної мови необхідно виконати певні етапи аналізу мовних структур, котрі неявно містяться в будові операторів M та M^{-1} . Такими етапами, зокрема, є: аналіз знакових систем мови, граматичний аналіз мовних структур, синтаксичний аналіз, семантичний аналіз, статистичний аналіз, контекстний аналіз і так далі.

Створення ефективних автоматичних процедур, орієнтованих на виконання відзначених різновидів аналізу, є досить складним науково-технічним завданням. На даний момент у світі вже існують програмні засоби, які певною мірою спроможні виконувати такі операції. Найпоширенішими серед них є пошукові системи, зокрема Інтернетівські, які можуть виконувати цілу низку природномовних функцій. Тим не менше, задоволення у користувачів існуючих засоби не викликають. Навпаки, всі добре знають, як важко там знайти потрібне знання, наскільки “шумлячими” є пошукові системи Інтернету. Недарма останніми роками активно просувається програма так званого Semantic Web⁸, покликана розробити інтелектуальні засоби екстракції знань з глобальної мережі.

Наш досвід переконує, що такі засоби і, відповідно, технології повинні мати комплексний характер, адже годі сподіватися на те, що одна схема спроможна охопити все багатовиддя когнітивних ситуацій, котрі виникають при інтелектуальному опрацюванні текстів на предмет екстракції з них певних знань. При розробці лінгвістичних технологій нового покоління постає завдання проведення фундаментальних досліджень системних зв'язків у тріаді “інформація — мова — інтелект”. Провідну роль тут відіграє встановлення принципів формального моделювання мовної системи.

З цією метою зробимо декілька методологічних зауважень щодо принципів моделювання мовної субстанції.

Виходячи з того, що власними об'єктами мови виступають певні психофізичні стани та процеси, які відбуваються в мовнорозумовому апараті людини, а усна та писемна її репрезентації слугують елементами інфраструктури мовного процесу, з'ясуємо, якою є їхня роль у процесах моделювання мови.

Очевидно, що мовнорозумовий процес сам по собі є інтегрованим, тобто таким, що містить як мовний, так і ментальний компоненти. В мовнорозумовому апараті він має представлення у вигляді динамічної системи взаємопов'язаних рефлексів різного характеру, про зміст яких можна довідатися, наприклад, з книги В.М.Бехтерева,

⁸ World Wide Web Consortium (W3C): <http://www.w3c.org/>

⁹ Бехтерев В.М. Объективная психология. — М.: Наука, 1991. — 480 с.

котра досі не втратила своєї актуальності. З огляду на викладене як усна, так і письмова форми мови відіграють ролі моделей мовнорозумових процесів і водночас — комунікативного для них середовища.

З іншого боку, психофізичні стани та процеси (а серед них і мовнорозумові), як правило, не є повною мірою досяжними для безпосереднього спостереження, а тим паче — об'єктивної фіксації. Отже, усна та письмова форми мови, фактично, слугують репрезентантами спостережуваних компонентів станів мовних об'єктів та процесів, що відбуваються в мовнорозумовому апараті. Як такі, вони, певне, і використовуються у ролі основних об'єктів концептуального моделювання мови.

Отже, вихідними положеннями розробки формальної моделі мови, орієнтованої на інтелектуальне її опрацювання, є такі:

— будь-яка мовна одиниця у контексті або в мовному потоці перебуває у певному — семантичному — стані;

— при розгляді формальних аспектів семантики ми виходимо з існування відповідності між мовною одиницею та її семантичним станом: $\psi: X \rightarrow \psi(X)$, де X — певна одиниця мови; ψ — відповідність між X та $\psi(X)$ — формальним об'єктом, що репрезентує семантичний стан одиниці X , який має своїми детермінантами елементи засобів матеріального вираження семантики;

— процес розуміння мови з цієї позиції виглядає як редукція апріорного розподілу станів мовних одиниць за елементами, що специфікують систему станів в цілому, до певного єдиного стану, притаманного саме тій мовній ситуації, яка є об'єктом мовної обробки у даний момент;

— на лексичному рівні це є редукцією апріорного розподілу лексем за тими ознаками граматичної та лексичної семантики, які притаманні суб'єктивному лексикону реципієнта, до певного єдиного граматичного та лексичного значення, притаманного саме тому контексту, який перебуває у полі уваги реципієнта і підлягає в даний момент процесу його індивідуальної мовної обробки.

Концептуальні засади теорії семантичних станів та певні практичні наслідки цієї теорії викладено у наших працях¹⁰; викладові цього підходу присвячено розділ 5 цієї монографії.

Формальні детермінанти знань, що містяться у природномовних текстах, знаходять свій зв'язок з параметрами мовної системи через

¹⁰ В.А.Широков. Феноменологія лексикографічних систем. Розділ 6. — К.: Наукова думка, 2004; В.А.Широков. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії. “Мовознавство”, 2005, №№ 3-4; Широков В.А. Семантические состояния языковых единиц. ББК 81.1Труды международной конференции “MegaLing'2005. Прикладная лингвистика в поиске новых путей”/ Отв. ред. В.П. Захаров, С.С. Дикарева. — СПб.: Издательство “Осьпів”, 2005. — 180 с. ISBN 5-98883-013-7 с. 147 — 162.

будову операторів M та M^{-1} та специфікацію семантичних станів $\psi : X \rightarrow \psi(X)$.

Тим часом, параметри, що специфікують “матричні елементи” операторів M та M^{-1} , зазвичай невідомі, та навіть більше — досі не дослідженою залишається формальна схема, за допомогою якої вони можуть бути встановлені. Отже, на часі проведення широкомасштабних експериментів над природномовними текстами, за допомогою яких зазначені параметри (або хоча б якась їх частина) були б встановлені емпіричними методами, що надасть матеріал до узагальнень та подальших досліджень. Саме на цьому шляху можуть бути розроблені ефективні інформаційно-комп’ютерні технології опрацювання знань, які становитимуть технологічну основу економіки знань — нової фази постіндустріального суспільства.

Таким чином, на даний момент сформувалася й актуалізувалася суспільна потреба у постановці на твердий експериментальний ґрунт питань когнітивної лінгвістики та її підрозділів — когнітивної семантики та когнітивної лексикографії, для чого потрібні великі, різнопланові, багатогалузеві та семантично марковані лексикографічні системи, спроможні представити у впорядкованому й максимально повному вигляді репертуар мовних одиниць різних рівнів, а також відношень та зв’язків між ними. Як наслідок, впливає й необхідність мати універсальну концептуальну схему, спроможну охопити весь комплекс явищ, всю феноменологію лексикографічного опису мовної системи. Саме ця настанова послугувала стимулом до написання цієї книги. Своїм завданням ми поставили подати виклад лексикографічної ділянки опису мовної системи не тільки (і не стільки) у її традиційному розумінні, а насамперед з огляду на нові когнітивні процеси та нові функції словникових систем у сучасних засобах опрацювання інформації. При цьому нашим прагненням було викласти фундаментальні засади неklasичної теорії лексикографії, подавши водночас і максимально можливу кількість її застосувань (у тому числі і для традиційних лексикографічних об’єктів) з метою продемонструвати працездатність та корисність розвинутої теорії на змістових, подекуди вельми складних прикладах. Зауважимо, що одним з практичних наслідків застосування на практиці нашої лексикографічної концепції стало створення Національної словникової бази України, яку рішенням Уряду України¹¹ віднесено до числа наукових об’єктів, що становлять національне надбання України; її основу складає серія фундаментальних академічних словників нового покоління — “Словники України”.

¹¹ Розпорядження Кабінету Міністрів України від 11.02.2004 р. № 73-р “Про віднесення наукових об’єктів до таких, що становлять національне надбання”.

Монографія складається з шести розділів.

Перший розділ “Теоретичні засади фундаментальної лексикографії”, починаючись з огляду традиційних засад лексикографії, функцій, структури, типології словників, ролі та місця лексикографії у мовознавчій науці, завершується викладом теорії лексикографічних систем, яка надає далекосяжне узагальнення поняття словника і водночас відкриває нову його філософію, а також і прагматику створення та застосування словникових систем у сучасній практиці.

У другому розділі “Електронні граматичні словники” викладено концептуальну модель та комп’ютерну реалізацію явищ словозміни для української та російської мов, а також побудову на цій основі інструментальних комп’ютерних комплексів для створення та ведення електронних граматичних словників.

Третій розділ присвячено аналізу структури та побудови лексикографічної системи академічного Словника української мови, а на цій базі — розробці інструментальної комп’ютерної системи фундаментальної тлумачної лексикографії, яку реалізовано в Українському мовно-інформаційному фонді при укладанні нового 20-того Словника української мови.

У четвертому розділі “Інформаційно-лексикографічне моделювання інтегрованих словникових систем” викладається повна схема реалізації архітектури Інтегрованої лексикографічної системи “Словники України”.

У п’ятому розділі “Семантичні стани мовних одиниць та їх роль у моделюванні мови” вводиться поняття семантичного стану мовної одиниці та будується нарис формального опису семантичних станів. Із застосуванням апарату теорії нечітких множин демонструється застосування семантичних станів для опису складних мовних явищ.

Шостий розділ присвячено проблемі створення багатомовних лексикографічних систем з метою їх застосування в контурах інтелектуальної мовної обробки, зокрема при машинному перекладі. Аналізується роль синонімії у цих процесах та викладається схема побудови універсального інструментального комплексу багатомовної лексикографії. Зазначений виклад пропонується як в локальному варіанті, так і в рамках концепції віртуальних систем професійної взаємодії в лінгвістиці.

Автор дякує своїм співробітникам Т.О.Грязнухіній, О.М.Костишину, Т.П.Любченку, О.О.Погрібній, О.Г.Рабульцю, Н.М.Сухарині (Заїці), В.В.Чумаку, І.В.Шевченку та К.М.Якименку, у співпраці з якими одержано ряд результатів, які увійшли до цієї книги.

Автор також виносить щире подяку науковим редакторам видання — Т.О.Грязнухіній та О.Г.Рабульцю, а також рецензентам академіку НАН України В.М.Русанівському та члену-кореспонденту НАН України О.В.Палагіну.