

Аналізуючи розвиток комп'ютерної лексикографії, у ній можна виділити три головні напрями:

1. Автоматизована побудова та укладання традиційних словників, реєстрів, конкордансів, лінгвістичних корпусів тощо.

2. Розробка комп'ютерних словників як елементів інформаційно-пошукових підсистем та інших засобів лінгвістичного забезпечення автоматизованих систем.

3. Створення лексикографічних систем, як принципово нового типу мовно-інформаційних об'єктів, здатних до автономної взаємодії з відчуженими від людини текстами¹⁰.

З метою встановлення своєрідного "містка" між класичною та некласичною ділянками лексикографії звернімося до визначення поняття словника та опису основних елементів його структури. Це, на нашу думку, допоможе нам у подальшому, а саме — у розкритті тих онтологічних механізмів, котрі спонукають мовну субстанцію набувати словникової форми.

1.3. Означення, функції та загальна структура словника

Філологічна наука за багатовікову історію свого існування і розвитку виробила системні підходи до означення поняття словника. У вузькому розумінні словник (лексикографічний твір) — це сукупність мовних одиниць, розташованих у певному порядку, де розкрито їхнє значення, подано про них різні відомості чи переклад іншою мовою або вміщено інформацію про предмети, явища та факти, які вони позначають.

Більш широке означення словника, яке належить бельгійському лексикографу К. К. Бергу (цитуюмо за працею¹¹), формулюється так: словник — це систематизований перелік соціалізованих (тобто тих, що стали стабільними в системі мови) мовних форм, взятих зі звичайного мовлення певної мовної спільноти і описаних укладачем так, що підготований читач розуміє зміст кожної окремо взятої мовної форми і отримує інформацію про істотні факти, пов'язані з функціонуванням цієї форми в даній мовній спільноті.

Якщо ж словник розуміти як підсумок дослідницької роботи над якоюсь проблемою, як розв'язок певної лінгвістичної проблеми чи

¹⁰ Широков В. А. Інформаційна теорія лексикографічних систем. — К.: Довіра, 1998. — 331 с.

¹¹ Zgusta L. Manual of Lexicography. Praha: Nakl. Ceskosl. akad. ved., 1971.

як інструмент розв'язання нових завдань, то поняття “словник” розширюється у напрямі позначення систематизованої сукупності знань про певну проблему¹² і набуває цілком визначених конотацій до когнітологічних структур типу моделей знань.

Теоретична лексикографія подає означення словника як абстрактного мовно-інформаційного об'єкта, визначальними рисами якого слугує певна низка системно-структурних ознак.

Передусім — це членоване розміщення матеріалу, оскільки основною композиційною та комунікативною одиницею словника¹³ є відносно самостійний відрізок тексту, що називається словниковою статтею; множина словникових статей і складає основу словника.

У свою чергу, кожна словникова стаття є двокомпонентною — в ній виділяється реєстрова (“ліва”) та інтерпретаційна (“права”) частини. “Ліва” частина — це, зазвичай, будь-яка одиниця мови, що є об'єктом лексикографування і вноситься до словника. Вона називається ще лексикографічною одиницею. Множина реєстрових одиниць словника складає його реєстр. Структура і зміст “правої” частини словникової статті залежить від типу словника; тут подається лінгвістичний опис відповідної реєстрової одиниці (її значення, переклад іншими мовами, наголос, лексичні, граматичні характеристики тощо) або міститься різноманітна інформація про предмети, які вона позначає. На склад реєстру та структуру словникової статті, окрім того, дуже впливає кінцеве призначення словника¹². Зазначимо, що терміни “ліва” та “права” частини словникової статті є певною мірою умовними, оскільки у деяких типах словників спостерігається “чергування” елементів структур реєстрової та інтерпретаційної частин. У теорії лексикографічних систем, постулюється наявність інваріантного принципу, який здійснює сегрегацію елементів структури.

Реєстрові одиниці у словнику пов'язані між собою численними структурно-семантичними зв'язками і в своїй сукупності творять певну систему, яка реалізує задуми та цілі його укладачів. Окрім того, словникові статті будь-якого словника повинні мати — в його межах — тотожні схеми опису однотипних елементів лексикографічних структур. Сукупність правил, прийомів та засобів опису реєстрових одиниць творять метамову словника.

¹² Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. — М.: Наука, 1981. — 367 с.

¹³ Денисов П. Н. Системность и связанность в лексике и система словарей // Проблематика определений терминов в словарях разных типов / Ред. кол. С.Г.Бархударов и др. — Л.: Наука, ЛО, 1976. — С.63-73. — С.68-69.

Важливою характеристикою словника є нормативність — вона, фактично, слугує (точніше, повинна слугувати) імперативною настановою для лексикографів. Зміст цієї настанови полягає в тому, що при описі кожної лексикографічної одиниці її слід певним чином співвідносити із конкретною формою існування (функціонування) мови, регламентуючи правила її мовленнєвого вживання. Доречно підкреслити, що укладач словника зазвичай лише описує та досліджує мовні процеси, визначає мовні закономірності, але аж ніяк не може зобов'язувати, вказувати, йти всупереч природному розвитку мови.

Текст словника є строго упорядкованим — реєстрові одиниці у кожному словнику обов'язково подають впорядкованими за якийсь критерієм (наприклад — за абеткою).

Нарешті — наявність допоміжного матеріалу. Будь-який словник окрім словникових статей містить, хоча б у мінімальній кількості, додаткову інформацію, що допомагає ним користуватись: передмову, критерії впорядкування та опис структури словникових статей, список скорочень, граматичні таблиці, джерела укладання тощо.

У структурному аспекті теоретична лексикографія розрізняє три взаємопов'язані рівні ієрархії: макро-, медіо- та мікроструктуру словника¹⁴.

До *макроструктури* відноситься все те, що визначає словник як самостійну систему з її внутрішніми зв'язками та багатоплановою організацією. У цьому плані теоретичне словникарство вирішує цілу низку проблем. Важливе місце серед них займає створення науково обґрунтованої типології словників; крім того макроструктурними завданнями є обґрунтування необхідності розроблення конкретного словника, вивчення та відбір джерел до його укладання, визначення загальної структури словника, опрацювання принципів його побудови та укладання, способи творення реєстру та визначення характеру лексики, яку належить включити в словник, визначення можливих перетворень лексичного складу мови (компресії, мінімізації) для укладання реєстрів словників різних типів, вироблення метамови словника, уніфікація композиції і апарату посилення однотипних словників для полегшення користувачам переходу від одного словника до іншого, принципи відбору ілюстра-

¹⁴ Агрикола Э. Микро-, медио- и макроструктура как содержательная основа словаря // Вопросы языкознания: — 1984. — №2. — С. 72-87. Гринев С.В. Введение в терминологическую лексикографию. — М.: МГУ, 1986. — 106 с.

тивного матеріалу, визначення критеріїв упорядкування реєстрових одиниць, відображення у словнику статистичних даних тощо.

До основних проблем *медіоструктурного* рівня словника відносяться побудова та експлікація різноманітних лінгвістичних відношень між словниковими одиницями. Вони втілюються в об'єднанні реєстрових одиниць у поля та групи на основі морфологічних, семантичних, тематичних, асоціативних та інших ознак (прикладом структр медіорівня є, зокрема, відношення словотвору, яке у словниках певного типу набуває вигляду словотвірних гнізд).

На *мікроструктурному* рівні вирішуються проблеми словникової статті: її структура, форми та способи розкриття семантики реєстрових одиниць, ієрархія їхніх значень тощо. Теоретичною лексикографією встановлено такі вимоги до лексикографічного опису як *стандартність* (єдиний протягом цілого словника опис однотипних явищ), *економність* (надання переваги — за однакових інших обставин — коротшому опису перед довшим, *простота* (застосування при лексикографічному описі максимально простих та зрозумілих синтаксичних конструкцій і слів, в лексичному значенні яких нема непотрібної двозначності і зайвої розпливчастості), *повнота* (прагнення вичерпного опису всіх важливих значень і способів вживання реєстрових одиниць).

Ідеалом традиційної лексикографії вважається універсальний словник, під яким розумітимемо якомога повніше зібрання мовних одиниць конкретної мови, де про кожен з них уміщено всеохоплюючу інформацію у всіх її можливих аспектах. Це означає, що універсальна словникова стаття повинна мати такі компоненти: *реєстрову одиницю*; *формальні характеристики* — фонетичні, граматичні, морфологічні, орфографічні, дериватологічні, синтаксичні, етимологічні, стилістичні тощо; *семантизацію*; *цитати з текстів*, що ілюструють ту чи іншу формальну або семантичну особливість лексикографічної одиниці; *зв'язок з іншими реєстровими одиницями у різних "координатах"* семантичного простору мови; різноманітні *довідки та відсилачі*; інформацію про предмет, факт, явище, відношення, який означає реєстрова одиниця.

Корисним поняттям для визначення структури словникових систем, а також для побудови їхніх класифікаційних схем слугує введене Ю. Н. Карауловим. [Караулов, 1981] поняття *лексикографічного параметра*. За його визначенням лексикографічний параметр — це деякий "квант" лінгвістичної інформації, що може мати самостійний інтерес для користувача, але, як правило, виступає в комбінації з іншими параметрами ("квантами") і знаходить своє специфічне вира-

ження у словниках; іншими словами — це особливе словникове відображення окремих структурних рис мови.

Для усунення неоднозначності у трактуванні лексикографічних параметрів надалі ті з них, що відносяться до словника в цілому (призначення словника, коло користувачів, спосіб використання, об'єм, кількість мов, входів тощо), називатимемо *словниковими параметрами*. Лексикографічними параметрами у власному значенні вважатимемо ті, що співвідносяться з лексикографічною одиницею (наприклад, орфографічний (правописний) параметр, графічна довжина слова, наголос, вимова (орфоепічний параметр), поділ на склади, частина мови тощо).

Лексикографічним параметрам притаманні певні властивості¹⁵, які мають універсальний характер і не залежать від типу словника, а їхні значення характеризуються певною національною специфікою (наприклад, рід іменників); у певних мовах певні параметри можуть бути і відсутніми. Кількість лексикографічних параметрів у словниках варіюється від одного до декількох десятків, причому ці коливання визначаються не лише цільовою установкою словника, але і його орієнтацією на того чи іншого користувача.

Лексикографічний параметр завжди відноситься до лексикографічної одиниці в цілому: це не склад, а поділ на склади, не афікс чи суфікс, а морфемне членування, не окреме словотворче значення, а комплекс словотворчих відношень або гніздо, не фонема і не звук, а вимова і т. д.

Як віддзеркалення структури всієї мови лексикографічні параметри різняться за своїм обсягом, подекуди вони здаються неспівмірними один з одним (наприклад, наголос та словотворче гніздо). Деякі параметри можна виділити досить легко (наприклад, довжина слова), визначення ж змісту інших вимагає попереднього, інколи значного, дослідження (наприклад, семантичний еквівалент слова або його етимологія).

Один і той же параметр можна задавати різними способами. Окрім того практично будь-який параметр залежно від кінцевого завдання словника розкривають з різним ступенем глибини та деталізації. Прагнення до деталізації параметрів проявляється в укладанні однопараметричних словників. Можна сказати, що кожен параметр, а в границі — кожне значення параметра, прагне стати

¹⁵ Городецкии Б. Ю. Проблемы и методы современной лексикографии // Новое в зарубежной лингвистике. — М.: Прогресс, 1983. — Вып. XIV. — С.5-23. Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. — М.: Наука, 1981. — 367 с.

окремим словником. Ця відцентрова тенденція, як відзначалося, є прямо протилежна доцентровій тенденції до універсальності, тобто об'єднання всіх лексикографічних параметрів в одному словнику. Ці різноспрямовані тенденції разом творять характерне прагнення лексикографії закріплювати результати мовних досліджень на всіх ділянках лінгвістики у формі словників. Слід зазначити, що збільшення глибини відображення одного параметра обов'язково веде до розширення опису його зв'язків з іншими параметрами мовної структури, до втягнення їх у сферу представлення цього параметра, тобто при досягненні певної глибини в розробці окремого параметра відцентрова тенденція перетворюється на свою протилежність. Багато параметрів виступають у синкретичній формі: акцентологічна інформація подається або в поєднанні з орфографічною, або прив'язується до орфоепічного параметру; складоподіл передбачає наявність даних про вимову слова і т. ін.

За своєю природою лексикографічні параметри діляться на дві групи. До одної з них входять власне мовні, тобто *структурогенні параметри* — наголос, орфографічний параметр тощо. Їхні значення є дискретними і можливість варіювання розповсюджується лише на способи задання, фіксації в словнику. Другу групу становлять параметри, зміст яких за необхідністю включає і екстралінгвістичний фактор — денотативний, історико-культурний, прагматичний, тобто моменти, вторинні у відношенні до мови, пов'язані не лише з самою мовою, але з її вивченням, тобто з мовознавством. Ці параметри вбачаються недискретними, характеризуються змінною глибиною наукового розкриття і відображають не власне структурні відношення, а відтворюють процеси — діахронічні, синтагматичні, взаємодії мов, інтерференції тощо. Називатимемо їх *лінгвістичними параметрами*. Розгляд сучасних словників різних мов не дав поки що причин говорити про наявність об'єктивних перепон на шляху побудови будь-якої сполучуваності параметрів, конструюванні своєрідного параметрично-лексикографічного простору. Цей висновок підтверджується й історичним розвитком лексикографії в напрямку нарощення числа параметрів в словнику. Характеризуючи параметри цих двох груп, відзначимо, що можна побудувати словники, які містять лише структурогенні параметри, але неможливе створення словників з одних лінгвістичних параметрів — останні обов'язково повинні доповнюватись і деякими структурогенними. З іншого боку, хоча самі параметри можуть сполучуватись у різних комбінаціях, далеко не всі способи їх задання та сукупність глибини виявляються сумісними один з одним. Таким чином,

одна з труднощів на шляху побудови універсального словника лежить у встановленні оптимального співвідношення між способами задання структурованих та ступенем глибини лінгвістичних параметрів у одному словнику. “Геометрична” інтуїція у цьому випадку твердить про знаходження змістових “гіперповерхонь” у параметричному просторі, які експлікують “правильне” співвідношення між різними лексикографічними параметрами.

Практично в усіх словниках вхід здійснюється за одним (рідше — за декількома) параметрами, але обов’язковою умовою при цьому є те, що ці параметри є лише структуровані і не можуть бути лінгвістичними. Більше того, число вхідних параметрів виявляється досить обмеженим, тобто далеко не всі структуровані параметри можуть реально використовуватись як вхідні. В універсальному словнику, де теоретично мусять поєднуватися всі лексикографічні параметри, бажано мати декілька входів, і не лише від основних структурованих параметрів, але й від ряду лінгвістичних. Зрозуміло, що ця проблема дуже важко розв’язати методами традиційної лексикографії.

Лексикографічні параметри закріплюють формальне вираження своїх значень, як правило, в системі словникових ремарок, хоча є параметри (наприклад, дефініція, орфографія), що спеціальних ремарок не мають. Ремарки однакового порядку можуть бути і значеннями одного параметра, допускаючи його більшу чи меншу деталізацію, наприклад, ремарки стилістичні, але можуть бути і різними параметрами, як у випадку системних відношень (синоніми, антоніми тощо).

В існуючих класифікаціях та методах типології словників встановлено інваріантні концептуальні поля, що відіграють роль своєрідних “систем координат”, кожна з яких параметризує певну групу властивостей мови¹⁶.

Так, наприклад, у лінгвістичній системі координат визначаються аспекти аналізу мови, системні ознаки форми та змісту мовних одиниць, види семантичної інформації про мову. *Психологічна сис-*

¹⁶ *Казакевич О. А.* Автоматизация лексикографических работ: Автоматические словари. // НТИ. Сер. 2. 1985. — №9. — С. 25-29. *Кустова Г.И., Падучева Е.В., Рахилина Е.В., Родина Р.И., Филипенко М.В., Якубова Н.М.; Янко Т.Е.* Словарь как лексическая база данных: об экспертной системе “Лексикограф”. // НТИ. Сер. 2. — 1993. — №11. — С. 18-20. Лінгвістическіе ісследованія. Воіросы лексикології, лексикографії і прикладної лінгвістики. / Отв. ред. Р.П.Рогожнікова. — М., 1976. — 232 с. *Пецак М.М., Клименко Н.Ф., Картиловская Е.А., Шкуров В.А., Цимбалюк И.В.* Украинский семантический словарь: Проспект. — К.: Наук. думка, 1990. — 264 с. Теорія і практика сучасної лексикографії: Сб. научних трудов. / Отв. ред. Р.П.Рогожнікова. — М.: Наука, Ин-т русс. яз., 1984. — 183 с.

тема відповідає за врахування особливостей сприйняття людиною істотних характеристик словникової системи (останнім часом слід говорити про людиномашинний аспект цього сприйняття). У *семіотичній* системі встановлюються критерії для створення метамов різних словників, визначаються способи і формальні засоби фіксації й репрезентації лексикографічної інформації. У *соціотехнічній* системі розв'язуються лінгвістичні, семіотичні, психологічні та технологічні словникові проблеми стосовно до конкретних часових, соціальних та виробничо-технологічних умов створення й використання словника і т.д.

Наступний крок на шляху побудови типології словників — встановлення таких характеристик, як вид словника (зокрема, словникові одиниці, обсяг словника, принципи його упорядкування тощо); метамовний апарат (інвентар засобів опису одиниць лексикографування, правила побудови словникових статей, типи міжстатейних та міжсловникових відображень; інформаційний та пошуковий апарат словника; можливі його функції) тощо.

Для визначення існуючих та встановлення нових типів словників є декілька шляхів.

Традиційним є шлях побудови багатовимірних класифікаційних схем, що процедурно реалізується як часткова структурно-функціональна декомпозиція предметної галузі лексикографії. Типовим результатом цього напрямку є класифікаційна схема поняття “тип словника”, в якій виділяється вісім основних груп ознак. Комбінуючись, вони параметризують множину можливих типів словників: 1) принцип визначення змісту та значення одиниць словника; 2) спосіб організації словника; 3) обсяг словника; 4) ставлення до типу мовного спілкування; 5) відношення до мовної норми; 6) спосіб представлення лексичного значення слова; 7) відношення до синхронії та діахронії; 8) призначення (мета) словника.

У традиційній лексикографічній практиці сформувалися такі способи організації матеріалу в словнику: алфавітний (найпоширеніший), інверсійний (фактично теж алфавітний, тільки не від початку реєстрових одиниць, а від кінця), словотвірний, тематичний, ідеографічний, частотний, семонімічний. Н.Ф. Колесников¹⁷ об'єднує словники синонімів, антонімів, паронімів і омонімів у групу семонімічних словників, в яких “зібрані і тлумачаться не окремі слова, як у звичайних тлумачних словниках, а два слова і більше, при об'єднанні яких враховуються відношення між їх звучанням і

¹⁷ Колесников Н.Ф. Семонимические словари. — Ростов, 1981. — С. 5.

(чи) значенням¹⁸) — вони також фактично зводяться до алфавітного; таким чином, навіть якщо паперовий словник, представляючи певні системні зв'язки між лексикографованими одиницями, і уможливорює їх експліцитне представлення (наприклад, подаючи тематичні чи ідеографічні групи, словотвірні чи синонімічні гнізда та под.), він підпорядковується алфавітному впорядкуванню, раз і назавжди зафіксованому поліграфічним методом.

Традиційний словник обмежений щодо обсягу — реальної фізичної можливості подання лексикографічного матеріалу. Сучасне поліграфічне виробництво все ж таки є на порядки дорожчим у підготовці багатотомного видання, аніж випуск, скажімо, лазерного диску, на якому може зберігатися понад 300 тисяч сторінок звичайного тексту, не кажучи уже про інші види інформації (відеофрагменти, звуковий супровід та ін.). Тому очевидно, що прийнятий поділ традиційних словників за обсягом — великий (повний, інтегральний), малий (вибірковий, короткий), середній і тезаурускарбниця¹⁸ — не є релевантним для комп'ютерних словників та лексикографічних баз даних, хоча й за традицією вживається у їхніх назвах.

Згадуючи ці об'єкти у зв'язку з тим, що комп'ютерне лексикографічне моделювання мовних систем з суто філологічної проблеми перетворилося на нагальну потребу створення лінгвістичних компонентів програмного забезпечення, наголосимо на безперечній актуальності відтворення у лінгвістичному забезпеченні якомога більшого числа ефектів природної мови, причому максимально гнучким та універсальним способом. Цим пояснюється справжній бум у створенні комп'ютерних словників, оскільки словник здається тим компактним і добре впорядкованим середовищем, яке відносно легко піддається алгоритмізації і може бути ефективно використаний у лінгвістичних комп'ютерних компонентах. У зв'язку з такою активністю в галузі лінгвотехнології актуалізувалася потреба визначення кола словникових систем, необхідних для комп'ютерних застосувань. А це зумовлює необхідність повернення до проблем класифікації і типології словників уже на новому рівні, а саме такому, де і традиційні, і комп'ютерні словникові системи розглядаються з єдиного методологічного погляду. Зазначений підхід має не тільки теоретичне, а й технологічне значення, оскільки дозволяє встановити відповідність між традиційними і комп'ю-

¹⁸ *Роменская В.Ф.* О классификационной схеме понятия "тип словаря" в информационном тезаурусе // Структ. и приклад. лингвистика. — Л., 1978. — Вып.1. — С.181-187.

терними словниковими системами і в такий спосіб скористатися всім надбанням традиційної лексикографії як у власне комп'ютерному словникобудуванні, так і взагалі при побудові лінгвістичного забезпечення інформаційних систем. Такий шлях до створення класифікаційних схем словників, на наше переконання, відкриває розвинена у працях останнього десятиріччя теорія лексикографічних систем¹⁹, до викладу якої ми зараз переходимо і згідно з якою словники ідентифікуються як частинні випадки лексикографічних систем, для яких виділяються такі класифікаційні ознаки:

- типи лексикографічних ефектів;
- класи елементарних інформаційних одиниць (відносно кожного лексикографічного ефекта);
- структури елементарних лексикографічних систем (відносно кожного лексикографічного ефекта);
- процеси рекурсивної редукції лексикографічних систем;
- процеси інтеграції лексикографічних систем та будова відповідних лексикографічних середовищ;
- архітектурні реалізації лексикографічних систем та середовищ.

Навіть самі назви і репертуар перелічених класифікаційних чинників незвичні для класичної лексикографії. Тому одне із завдань нашої праці — довести і продемонструвати на змістових прикладах, що наведений інструментарій репрезентує універсальну класифікаційну схему, яка є корисною не тільки у застосуванні до будь-яких типів традиційних та комп'ютерних словників, але й охоплює широке коло інформаційно-лінгвістичних об'єктів (у тому числі й таких як граматичні та логіко-лінгвістичні числення). Отже, зараз ми перейдемо до викладу теорії лексикографічних систем, метою чого є не тільки побудова корисної формалізованої схеми лексикографічного опису мови, але й — і насамперед — з'ясування того онтологічного принципу, який спонукає мовну субстанцію "самоорганізовуватися" у словникову (чи словникоподібну) форму. Повчально, що для цього нам доведеться вийти з кола суто мовних явищ і проаналізувати набагато ширший та універсальніший феноменологічний простір.

¹⁹ В.А.Широков. Інформаційна теорія лексикографічних систем.; Широков В.А. Строе-ние лексикографических систем. //Математические машины и системы. — К., № 2, 1999, С.83—104.; В.А.Широков, О.Г.Рабулецъ. Формалізація в галузі лінгвістики. 3б. пр. Актуальні проблеми української лінгвістики. 2002, Вип. V. С.3-28.

¹⁹ В.А.Широков. Феноменологія лексикографічних систем.

1.4. Теорія лексикографічних систем

1.4.1. Лексикографічний ефект в інформаційних системах

Процеси лексикографування, як різновиду інтелектуальної діяльності, та феноменологія словників, що виступають результатами цієї діяльності, не є постійними у часі величинами — вони еволюціонують відповідно до внутрішнього розвитку лінгвістичної науки та потреб практики. У певні історичні періоди саме зовнішні по відношенню до власних завдань мовознавства чинники — як це вже неодноразово було в історії науки — перетворюються на головну рушійну силу, що визначає розвиток не тільки самої лексикографії, як окремої ділянки лінгвістики, але й мовознавчої науки в цілому. Ми переконані, що відзначений розвиток ще позначиться, і кардинально, також і на прогресі інформаційної науки — упевненість у цьому нам надає аналіз напрямку руху інформаційного суспільства.

Справді, з появою комп'ютерних технологій виник безпрецедентний в історії світової цивілізації феномен — спілкування людини з неживою істотою за допомогою і через посередництво природної людської мови. Як би скептично не ставитися до “розумових” потенцій обчислювальної машини, не можна відмахнутися від факту, що мовні реакції сучасного комп'ютера в деяких випадках уже неможливо відрізнити від реакцій людини, а зважаючи на прогрес лінгвотехнології протягом останніх двадцяти років, можна упевнено прогнозувати, що через деякий час комп'ютерами буде перебрано істотну частину мовної та лінгвістичної компетенції людини, що створить реальні передумови для побудови інформаційно-комп'ютерних систем на засадах природної мови.

Зазначена компетенція є важливою складовою частиною мовно-розумового апарату людини, а в цьому апараті, як уже досить твердо встановлено психо- та нейролінгвістикою, мовні структури є нерозривно пов'язаними зі структурами мислення. Означений зв'язок вбачається настільки суттєвим, що виправдовує визначення інтелекту як форми індивідуалізації систем, якій притаманний мовний статус. Отже, у розглядуваному контексті комп'ютерне моделювання мови виявляється конгеніальним, майже ідентичним до моделювання інтелекту.

Яку роль у цій справі відіграє словник? Як зауважив Р. Шенк²⁰, у людському мовнорозумовому апараті функціонують різні “слов-

²⁰ Шенк Р., Бирibaум Л., Мей Дж. К интеграции семантики и прагматики // Новое в зарубежной лингвистике : Вып 24. Компьютерная лингвистика. — М., 1988. — С. 33.

ники” (які навіть більше нагадують “енциклопедії”), а дослідження протягом останніх десяти років за методологією WordNet²¹, що ґрунтуються на принципах психолінгвістики, не лише підтверджують цю тезу, а й навіть проливають світло на конструкції структур суб’єктивного лексикону людини.

Та обставина, що існує принципова можливість моделювання людської мови на неживих об’єктах, якими є комп’ютери, нашою думкою, що в природі матерії також діють механізми, які мають з людською мовою спільні риси — отже феномени, які можуть бути кваліфіковані як вияви мови, не обов’язково мусять бути виключною прерогативою людських (і взагалі живих) істот. У науці клас явищ та об’єктів, що характеризуються як “мовні”, відзначається великою різноманітністю. По-перше, це сама природна мова, яка існує тільки у формі національних мов (їх зараз у світі налічується близько шести тисяч і взаємовідношення між ними набувають подекуди драматичного характеру²²). Існують також й інші природні семіотично-семантичні системи, які з певною метафоричністю визначаються як “мови” — наприклад, “мова” генетичного коду. Крім того, створено цілу низку артефактів мовної орієнтації: штучні мови типу есперанто, що “імітують” певні природні; формальні алгебро-алгоритмічні конструкції, які одержали назву формальних мов і в яких останнім часом усе помітніше зближення з природними мовами (в математичній лінгвістиці навіть існує термін “майже природна мова”). Про поширеність терміну “мова” в інформаційній науці свідчить терміносистема інформатики, яка містить такі поняття: мова програмування, інформаційно-пошукова мова, мова класифікаторів, мова запитів, мова опису даних, мова маніпулювання даними, мова опису предметної галузі, цілка низка так званих “мов розмічування” (SGML — Standard General MarcUp Language, HTML — HyperText MarcUp Language, XML — eXtended MarcUp Language, VRML — Virtual Reality MarcUp Language) і так далі. У 60—70-і роки минулого століття сформувалася ціла галузь інформатики — лінгвістичне забезпечення інформаційних систем.

Виняткову роль у культурі й цивілізації відіграють штучні інтерпретації природних мов — такими, зокрема, виступають їхні писемні варіанти (інколи, забуваючи про штучне походження письма, писемний варіант, який є, по суті, не більше ніж моделлю

²¹ <http://cogsci.princeton.edu/~wn/>.

²² Дьячков М. В. Миноритарные языки в полиэтнических (многонациональных) государствах. — М., 1996. — 116 с.

мови, називають просто — мовою). Усі штучні мовні конструкції адаптують ті чи інші ознаки і властивості природних мов, а їх зближення з останніми містить ознаки тих національних мов, носіями яких є автори зазначених штучних конструкцій (зокрема, природномовна орієнтація інформаційно-комп'ютерних мов поки що зосереджується у колі так званих світових мов, переважно англійської).

У всіх названих “мовних” систем є спільні риси; вони мають основоположний, фундаментальний характер, який — ми переконані — безпосередньо пов'язаний із визначенням лексикографічних процесів і структур, причому таким визначенням, що враховує нові їх аспекти, про які йшлося вище. Для осмислення цих рис і побудови працездатної концептуальної схеми нам необхідне таке узагальнене уявлення про мову, яке можна було б застосувати до будь-якої з перелічених “мовних” систем, адже в сучасному інформаційному середовищі мова перестає бути виключною прерогативою людини, принаймні на “технологічному” рівні.

Що говорить про це лінгвістична наука? Вона твердить (висловлюючись як і кожна фундаментальна наука про свій предмет), що питання про визначення мови не просте, їх існує багато й вони різні²³ — з огляду на ті чи інші аспекти цього багатогранного феномену. Незважаючи на їхню розмаїтість, узагальнюючи їхні істотні риси, доходимо висновку, що більшість із них виявляється варіаціями на тему, запропоновану Чейфом²⁴: мова — це система, яка в досить складний спосіб здійснює зв'язок між світом звуків і світом значень.

Значимо, що існує багато інших дефініцій мови. Чимало з них ще з часів В. Гумбольдта ґрунтується на розмежуванні понять мови і мовлення. Після опублікування книги Ф. Соссюра “Курс загальної лінгвістики” це питання набуло широкого розголосу в лінгвістичних колах, причому діапазон думок коливався (і продовжує коливатися) у надзвичайно широких межах. Їх можна звести до трьох основних тверджень: 1) мовлення і мова протиставляються одне одному як цілком автономні об'єкти, що різняться між собою за сукупністю суттєвих ознак, унаслідок чого їх дослідження ведеться двома самостійними галузями науки — лінгвістикою мови та лінгвістикою мовлення; 2) мова і мовлення становлять єдиний об'єкт

²³ Є ще й інший погляд, згідно з яким формальне визначити предмет науки, залишаючись у її межах, взагалі неможливо.

²⁴ Чейф У. Значение и структура языка. — М., 1975. — 432 с.

лінгвістики, а тлумачення відмінностей між ними покладається на методологію та зміст цієї комплексної науки; 3) між мовою і мовленням узагалі немає жодної різниці.

Л. В. Щерба²⁵, як відомо, виділяв три основні аспекти мовних явищ: під першим він розумів *мовленнєву діяльність*, під другим — словники і граматики, які виводяться на підставі актів мовлення і розуміння, що існували в певний історичний час серед людей певних груп, тобто *мовні системи*, під третім — сукупність того, що говорили або розуміли такі групи — *мовний матеріал*. Він підкреслював, що мовна діяльність зумовлюється складним мовним апаратом людини або психофізіологічною *мовленнєвою організацією* індивіда, яка: а) не може дорівнювати сумі мовленнєвого досвіду й повинна бути своєрідним його опрацюванням; б) може бути тільки психофізіологічною; в) як і мовленнєва діяльність, є соціальним продуктом; г) слугує індивідуальним виявом мовної системи, що виводиться з мовного матеріалу; про характер цієї організації можна судити тільки на підставі мовленнєвої діяльності індивіда. Л. В. Щерба розмежовував поняття механізму (*мовленнєвої організації* людини) і процесу (*мовленнєвої діяльності*); процесу і його продукту. Мовленнєву організацію він розглядає як єдність процесу й продукту; останній виступає як індивідуальна система концептів і стратегій використання індивідами у процесах мовлення й розуміння, яка і позначається як *мова*.

Ми не дискутуватимемо тут про відносну правильність тих чи інших поглядів, оскільки всі вони містять ті чи інші характерні ознаки того феномену, для якого використовуємо спільну назву “мова”, і варіюються в різних комбінаціях у великому масиві лінгвістичних праць загального характеру. Ми опускаємо тут їхній виклад, оскільки вважаємо, що в їхній основі — явно або неявно — також лежить відповідність між світом звуків і світом значень. Не беремо до уваги також і визначень типу “мова — душа народу” та аналогічних, оскільки їх не вдається дослідити науковими методами.

Зазначимо, що для побудови продуктивної схеми (навіть мінімально формалізованої) використати визначення, про які йшлося, важко. Якщо “світ звуків” ми ще можемо якось “локалізувати”, то зі “світом значень” справа набагато складніша. Справді, де цей світ зосереджений? Як його дістатися? Як ним “оперувати”? Взагалі, який смисл має твердження про існування “світу значень”? Як зі “світом значень” співвідносяться “світ змістів”, “світ зображень” і

²⁵ Щерба Л. В. Языковая система и речевая деятельность. — Л., 1974. — 428 с.

багато інших “світів”? Отже, поєднання в одному визначенні понять, що дуже різняться між собою за ступенем абстракції (“світу звуків” зі “світом значень”), викликає відчуття логічного розриву в цій дефініції і ставить більше запитань, ніж дає відповідей.

Проте і викладену, й інші спроби означення мови вважаємо корисними, оскільки кожна з них дає матеріал для узагальнення та абстрагування істотних рис такого універсального явища, яким є мова. Проаналізувавши їх і придивившись (“прислухавшись”) до будь-якого явища, яке ми інтуїтивно ідентифікуємо як “мову” з огляду на те, “що вона робить” і “як вона працює”, можемо зробити такий висновок: мова становить певний “інструмент” — своєрідну “машину”, яка забезпечує перетворення “форми” на “зміст” і навпаки. Але не довільної “форми”, а “форми” лінійних послідовностей певних дискретних об’єктів (звуків і звукових комплексів, знаків і знакових комплексів тощо).

Це твердження, не відрізняючись новизною²⁶, дає привід для заглиблення у феноменологію відношення “форма — зміст” (ВФЗ) і з’ясування таких його деталей, які б допомогли нам у абстрагуванні суттєвих рис мови. При цьому ми вважаємо, що в онтологічному вимірі ВФЗ не є апіорною якісною іманентною об’єкта як “речі в собі”, а такою його властивістю, що *розкривається* перед суб’єктом (“дається”) у процесі його взаємодії з об’єктом. Ми прагнемо до побудови формалізованого опису ВФЗ, адаптованого до створення спеціальної моделі даних, у якій були б представлені ефективні механізми рецепції істотних властивостей мови з технологічною орієнтацією на підтримку процесу створення словників та інших лінгвістичних продуктів. Для проведення аналізу деталей розгортання ВФЗ розглянемо діаграму, де символічно зображено процес сприйняття якогось об’єкта певним суб’єктом:

$$S : D \longrightarrow V(D). \quad (1.1)$$

Тут літерою D позначено “щось” з реального (або уявного) світу, що виступає в ролі об’єкта процесу сприйняття (спостереження, вивчення, уваги, переживання...) з боку певного S , яке ми вважаємо суб’єктом даного процесу; через $V(D)$ позначаємо результат цього процесу. Зауважимо, що в ролі S може виступати людина або сконструйований людиною прилад, або людино-машинна система, або

²⁶ Різні аспекти відношення “форма — зміст” активно експлуатуються в мовознавстві з часів В. Гумбольдта, а потім і Ф. Соссюра; у попередньому розділі ми також виклали установлені лінгвістичні погляди на це відношення.

будь-що інше, наділене властивостями сприйняття й відчуття (“відображення”); S може бути і “колективним суб’єктом” — групою людей, соціальною спільнотою, етносом, нацією, народом, сукупністю народів або навіть людством у цілому.

Внаслідок фізичної, психічної, інтелектуальної та іншої обмеженості суб’єкта S уся сукупність властивостей об’єкта D для його сприйняття розподіляється на дві, не дуже чіткі, неоднозначні, мінливі й не чітко відокремлювані частини. До першої з них залучаємо ті властивості D , які безпосередніше сприймаються “перцептивно-сенсорним” апаратом S , — позначимо цю частину через $F(D)$ і трактуватимемо її як сукупність формальних властивостей D з погляду суб’єкта S , який сприймає. До другої частини включаємо властивості D , які безпосередньо не сприймаються перцептивно-сенсорним апаратом S , а відбиваються в ньому опосередковано. Позначимо цю частину через $C(D)$ і розглядатимемо її як сукупність змістових властивостей D — знову-таки з погляду сприйняття того самого суб’єкта S . У зв’язку з цим діаграма (1.1) набуває такого вигляду:

$$D \xrightarrow{S_F} F(D) \xrightarrow{H} C(D), \quad (1.2)$$

де символом S_F позначено дію “перцептивно-сенсорного” апарату суб’єкта S , результатом якої є сукупність формальних (з погляду S) властивостей D ; символом H позначено механізм, який здійснює зв’язок між формою та змістом і забезпечує цілісність сприйняття об’єкта D суб’єктом S (якщо йому справді вдається забезпечити зазначену цілісність). Водночас, припустивши існування механізму, який дозволяє перехід від D до $C(D)$, — позначимо цей механізм через S_C , — одержуємо таку трансформацію діаграми (1.2):

$$\begin{array}{ccc} & S_F & \\ D & \xrightarrow{\quad} & F(D) \\ S_C \downarrow & \nearrow H & \\ & C(D) & \end{array} \quad (1.3)$$

де, як бачимо, відбулася “декомпозиція” суб’єкта S на S_F та S_C , що реконструюють формальні та змістові властивості D відповідно.

Ми не схильні абсолютизувати описану схему. Між $F(D)$ і $C(D)$ немає чіткої межі, як її взагалі немає між формою та змістом. Також майже ніяк не деталізувалися властивості S , хоч із загальних міркувань і було здійснено його декомпозицію на S_F та S_C . Отже, цей підхід має всі ознаки феноменологічного, оскільки не спирається на припущення щодо можливої “конструкції” S і механізмів

його функціонування. З цих міркувань можна стверджувати, що викладена схема є досить загальною — в неї не закладено жодних “анзатців”, крім єдиного, специфічного для мови: $F(D)$ повинен мати лінійний характер, тобто зображатися лінійними послідовностями дискретних об’єктів, джерелом яких виступає певна скінченна множина.

З огляду на сказане навіть сама можливість існування такого явища як мова виступає наслідком фундаментальної властивості S “бути суб’єктом”; тобто таким, для котрого будь-що має свою зовнішню сторону (форму) і внутрішню (зміст).

Відношення між цими різними аспектами сприйняття, символічно зображені величинами S_F , S_C , H , відзначаються великим розмаїттям, джерелом якого є фундаментально притаманні (тобто такі, яких у принципі не можна позбавитися) властивості сприймаючого суб’єкта S : мінливість, нерегулярність, різноманітність, обмеженість, нечіткість тощо.

Важливий аспект ВФЗ пов’язаний з властивістю концентрації “уваги” суб’єкта S , на окремих фрагментах і “форми” і “змісту”. Це досягається “настроюванням” його перцептивно-сенсорної системи на деталі того, що він сприймає, внаслідок чого початкове ВФЗ модифікується: певні змістові його елементи набувають властивостей форми, а до формальних додаються нові, раніше не помічені деталі. До цього класу властивостей, зокрема, належать поняття внутрішньої і зовнішньої форми мовних одиниць.

Відзначимо ще одну рису процесу розгортання ВФЗ. Сучасна культурологія схильна трактувати її як символ постмодернізму, а саме — вплив суб’єкта на об’єкт, себто можливість зміни стану об’єкта D у процесі його сприйняття (спостереження, дослідження...) суб’єктом S . Річ у тім, що для того, щоб відбувся процес, символічно зображений діаграмою (1.3), у багатьох випадках необхідно якимось “активізувати” об’єкт D , щоб він “проявив” свої властивості, якими “цікавиться” S . В класичній науковій парадигмі вважалося, що таке збудження об’єкта можна зробити скільки завгодно малим і знехтувати ним, вважаючи що воно істотно не змінює стану об’єкта. Проте розвиток науки виявив, що насправді це, взагалі кажучи, не так.

Те, що було викладене вище, виявляється співзвучним із зовсім іншими явищами та їх формальними моделями, які були побудовані незалежно, з іншого приводу і для інших цілей. Йдеться про визначення інформації А.М.Колмогоровим²⁷.

²⁷ Колмогоров А. Н. Три подхода у определении понятия “количество информации”. В кн. Теория информации и теория алгоритмов. — М., Наука, 1987. — С.213-223.

Введення колмогорівської інформаційної міри має на меті уточнення поняття інформації, по-перше, без залучення ймовірнісного підходу, а по-друге, надання можливості застосування цієї міри до індивідуальних об'єктів.

Основна ідея підходу полягає в тому, що інформація про певний об'єкт вважається одержаною тоді, коли можливо відтворити (реконструювати) цей об'єкт (модель об'єкта) за деяким його скінченням описом (набором ознак). Побудова міри Колмогорова базується на таких фундаментальних поняттях, як алгоритм, машина Тьюринга, рекурсивна функція, і походить від ідей теорії складності обчислень (складності алгоритмів), яка, фактично, і є витком інтерпретації інформації як міри складності і структурованості систем, а також упевненості в універсальності такої характеристики як складність, оскільки будь-яка система, незалежно від її природи, характеризується певною складністю та має певну структуру, хоча б і тривіальну.

Відповідна, не переобтяжена деталями математична конструкція формулюється у такий спосіб.

Розглянемо певну зліченну множину $X = \{x\}$. Вважатимемо, що існує взаємнооднозначна відповідність між X та множиною D двійкових слів, які починаються з одиниці — іншими словами, нехай задано бієктивне відображення:

$$n: X \rightarrow D, \quad (1.4)$$

таке, що кожному $x \in X$ однозначно відповідає певний $d = n(x)$, $d \in D$, і навпаки. Вважатимемо, що:

1. $n(x)$ — загальнорекурсивна функція на D . Позначимо через $l(d)$ довжину двійкового слова $d \in D$, тобто число нулів та одиниць, яке у ньому міститься. Тоді $l(n(x)) = l(x) + C$, де C — певна константа.

2. існує однозначне відображення $\chi: X^2 = X \times X \rightarrow X$ таке, що для $\forall x \in X, y \in X \exists z \in X$, що $z = \chi(x, y) \equiv (x, y)$ і $n(z) = n(x, y)$ є загальнорекурсивною функцією від $n(x)$ та $n(y)$, причому

$$l(x, y) \leq C_x + l(y),$$

де константа C_x залежить лише від x .

Вважатимемо ізоморфізм (1.4) встановленим, так що множину X також розглядатимемо як множину двійкових слів.

Припустимо, що існує частково рекурсивна функція $\varphi(p, x)$, яка ставить у відповідність двійковому слову x двійкове слово y , причому $p, p \in D$ інтерпретується як алгоритм (або програма), яка "переробляє" x на y .

$$p: x \rightarrow y, \quad (1.5)$$

а φ репрезентує при цьому метод (мову) програмування. Без втрати загальності вважатимемо, що p для даного x задається певним двійковим словом.

Позначимо:

$$K_{\varphi}(y|x) = \begin{cases} \min_p l(p), \text{ якщо } \varphi(p, x) = y \\ \infty, \text{ якщо не існує скінченного } p \text{ такого, що } \varphi(p, x) = y. \end{cases} \quad (1.6)$$

Таким чином, $K_{\varphi}(y|x)$ є довжиною мінімальної програми p , яка переводить x в y при заданому методі програмування φ . Ця величина називається складністю y відносно x при даному φ . Звичайно, залежність величини складності від φ є недоліком описаного методу, але існує теорема²⁸, яка стверджує існування “найкращого” методу програмування A , такого, що для будь-якої частково рекурсивної функції φ справедлива нерівність:

$$K_A(y|x) \leq K_{\varphi}(y|x) + C_{\varphi}, \quad (1.7)$$

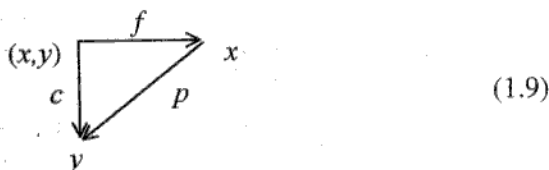
де стала C_{φ} залежить лише від φ і не залежить від x та y .

Величину $K_A(y) \equiv K_A(y|1)$, “віднормовану” відносно одиничного елемента $x = 1$, природно вважати складністю елемента y . При цьому кількість інформації в об'єкті x відносно об'єкта y визначається як різниця:

$$I_A(x|y) = K_A(y) - K_A(y|x). \quad (1.8)$$

Остання формула і визначає міру інформації — так звану алгоритмічну міру інформації Колмогорова.

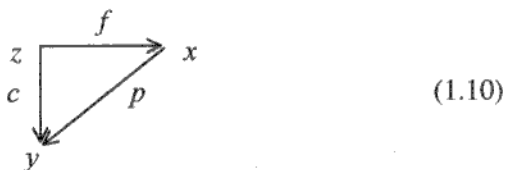
Формула (1.8) і весь описаний підхід визначення інформації через алгоритмічну складність можуть бути проінтерпретовані у децю інший спосіб. З цією метою збудуємо трикутну діаграму:



²⁸ А.Н.Колмогоров. Три подхода к определению понятия количества информации. В кн. Теория информации и теория алгоритмов. — М.: Наука, 1987. — С.220.

де x та $y \in X$; елемент $(x,y) \in X^2$, причому завдяки існуванню відображення $\chi: X^2 \rightarrow X$, $\chi(x,y) = z$, $z \in X$. У діаграмі (1.9) відображення f та c здійснюють проєкції елемента (x,y) на перший та другий співмножники, відповідно, причому справедливими залишаються формули (1.6)—(1.8) та інтерпретація складності та інформаційної міри.

Припустимо тепер, що об'єкт z , який досі у нас представляв образ декартового добутку (x,y) при відображенні χ , насправді є самостійним, до певної міри незалежним від x та y об'єктом зовнішнього світу. Це припущення дозволяє збудувати таку діаграму:



де $z \in Z$, Z — множина-джерело об'єктів z , $x \in X$, $y \in Y$. Тоді відображення f та c визначають певні інтерпретації об'єкта z , і, крім того, відображення p інтерпретує x через y . Природно припустити, що об'єкт x відбиває “формальні” властивості об'єкта z , а y — його “змістові” властивості, причому зв'язок між “формою” та “змістом” бере на себе p . Вимога мінімальності p тут є цілком зрозумілою, адже “витлумачення” форми (а його результат — це власне і є зміст!) не повинно містити “зайвих”, випадкових щодо x (а також і z) елементів. Побудована у такий спосіб конструкція, яка ґрунтується на трійці об'єктів (z, x, y) та трійці відображень (f, c, p) , що формують діаграму (1.10), допускає природну інтерпретацію як комплекс, який є носієм відношення “форма-зміст” (надалі вживатимемо для цього відношення абревіатуру ВФЗ).

Отже бачимо, що ВФЗ у такий свосередний спосіб виявляється “зашифрованим” в алгоритмічній теорії складності та інформації, а значить, це відношення є тісно пов'язаним з інформаційними процесами; навіть більше — воно, на наше переконання, є їхнім системотвірним відношенням. Тим часом воно є й основним системотвірним відношенням природної мови, що наполегливо і переконливо підтверджують відомі лінгвістичні теорії.

Отже, мовнорозумові об'єкти, процеси, конструкції та структури мають інформаційну природу, до якісного і кількісного аналізу якої цілком правомірне застосування формалізму Колмогорова. Очевидна “схожість”, спорідненість діаграми (1.3) з визначенням

інформації за Колмогоровим (див. формули (1.5—1.6)), а також подальші міркування, які привели до побудови діаграми (1.10)), наводять на думку про те, що в основі і одного і другого лежать ті самі закономірності. Сама форма представлення інформаційної міри свідчить про певний процес, результатом котрого стала генерація “алфавіту” — знакової системи представлення об’єкта. *Розгортання відображень* (f, c, p) у діаграмах (1.9)—(1.10), яким зіставляються елементи ВФЗ з діаграми (1.3), спонукає до зіставлення останніх зі складовими інформаційних процесів, які в теорії Колмогорова зводяться до математичних відношень (алгоритмів, рекурсивних функцій,...), визначених на дискретних множинах.

Описана ситуація здається настільки загальною, що дозволяє зробити висновок: в основі будь-якого процесу, явища, об’єкта, системи тощо лежить певна дискретна сукупність — назовемо її сукупністю *елементарних інформаційних одиниць* (надалі вживатимемо абревіатуру ЕІО), у визначенні якої ключову роль відіграє поняття *лексикографічного ефекту в інформаційних системах*. Цей ефект має семіотичне походження і суть його така.

Загальною ознакою всіх процесів інформаційного обміну є перетворення інформації з однієї форми в іншу, причому сучасні природничі теорії однозначно підтверджують принципово дискретний (“квантовий”) характер процесу взаємодії та обміну і, отже, принципову дискретизацію процесу опису дійсності. Зазначена дискретизація має принаймні одну спільну для всіх відомих процесів ознаку, на наш погляд, фундаментального характеру: спостерігаючи та узагальнюючи поведінку різних систем, доходимо висновку, що в процесі еволюції (динаміки, саморозвитку...) системи будь-якої природи в її структурі індукується певна підсистема відносно сталих дискретних сутностей (“підсистема порядку”), які відіграють роль її *елементарних інформаційних одиниць*, тому всі інші феномени системи являють собою певним способом організовані комбінації цих елементарних інформаційних одиниць.

Зазначена підсистема має властивості, споріднені з властивостями лексичної системи природної мови: вона “генерує” в своїй структурі щось на зразок тезаурусу і граматики з притаманними цим конструкціям знаковістю, значенням, змістом, поліморфністю, омонімією (ізоморфізмом), синонімією (гомоморфізмом), полісемією (багатозначністю), еліпсисом, метонімією, комбінованістю і т. д.; вона є носієм як “плану вираження”, так і “плану змісту”. Розгортання, взаємодія, взаємовпливи та взаємозумовленість обох зазначених планів у системі елементарних інформаційних одиниць

підлягають певним закономірностям, причому багатовікову контрарверзу натуралістів і конвенціоналістів ми схильні розв'язувати позитивно: існують приклади систем обох типів — як таких, у яких план вираження детермінований планом змісту, так і таких, зв'язок між якими має конвенціональний характер.

Сукупностям елементарних інформаційних одиниць властива “субстанційність”, як і будь-яким агрегаціям, спричиненим об'єктивними процесами (у даному випадку це лексикографічний ефект), унаслідок чого зазначені сукупності, як правило, мають відносну стабільність своїх характеристик, що забезпечує їхню локалізацію у відповідних областях системних параметрів.

Описана сукупність явищ становить зміст лексикографічного ефекту. Можна стверджувати, що при вивченні будь-яких предметних галузей фахівці фактично займаються дослідженням лексикографічних ефектів, які відбуваються в них або є для них характерними. Таким чином, лексикографічний ефект повинен розглядатися не лише з феноменального боку, а й з методологічного, оскільки він має певний “потенціал операціональності”, стимулюючи в процесі моделювання тих чи інших систем встановлення й визначення відповідних комплексів ЕЮ, враховуючи, репрезентуючи та конкретизуючи їхні властивості. У цій іпостасі концепція лексикографічного ефекту виступає як метод абстрагування даних.

У подальшому комплекс ЕЮ системи D , що розвинувся внаслідок лексикографічного ефекту (або ефектів) Q , позначатимемо через $I^Q(D)$ або $I(D)$, коли посилення на тип лексикографічного ефекту несуттєве. Система елементарних інформаційних одиниць, будучи носієм цілої низки властивостей, має певну структуру, що є їхнім виявом. Зокрема, в будь-якій системі ЕЮ завжди виділяється ядро — певна власна підсистема $I_0^Q(D) \subseteq I^Q(D)$ і визначається породжувальна процедура π :

$$\pi: I_0^Q(D) \rightarrow I^Q(D). \quad (1.11)$$

Трійку $(I^Q(D), I_0^Q(D), \pi)$ також ототожнюватимемо з системою ЕЮ і використовуватимемо це позначення поряд з $I^Q(D)$, $I(D)$, $I_0^Q(D)$, $I_0(D)$ як еквіваленти, вважаючи, що породжувальна процедура π визначена, відома і зрозуміла з контексту.

Підсумовуючи зміст викладеної феноменології, констатуємо, що процеси, аналогічні описаним в цьому підрозділі, відбуваються у всіх достатньо складно побудованих соціотехнічних системах і взагалі у системах будь-якої природи, в яких діють джерела, перетво-

рювачі та споживачі інформації і, отже, відбуваються аналоги перцептивно-сенсорних актів та мовнорозумових процесів. Це дає додаткове обґрунтування справедливості висновку про існування та універсальність інформаційного феномена — а саме лексикографічного ефекту в інформаційних системах з характеристиками *знаковості, значення, змісту й поліморфності*, що споріднюють його з природною мовою, якій також притаманні відзначені властивості²⁹. Вказані чинники підсилюють бажання побудови якомога формальніше визначених структур, які є носіями лексикографічних ефектів, та їхнього застосування до опису мовних феноменів.

1.4.2. Структура і архітектура лексикографічних систем

Базовим конструктивом побудов такого типу ми визначаємо так звані *лексикографічні системи* (надалі паралельно вживатимемо скорочення: Л-системи). Поняття Л-системи на наше переконання є фундаментальним, а його означення ґрунтується на викладеній феноменології лексикографічного ефекту. На наше переконання, лексикографічні системи уособлюють досить загальний тип формалізованих конструкцій у ряду таких як от: моделі даних, формальні граматики, формальні системи, канонічні числення у скінченних алфавітах (системи Поста) тощо.

Зазначимо, що з частинними випадками (або реалізаціями) лексикографічних систем у науці й техніці оперують дуже давно: ними виступає велика кількість різноманітних інформаційних систем, баз даних та знань, до кола яких потрапляють і всі традиційні словники та комп'ютерні словникові системи.

Якщо говорити про машинні словники, то вони спроможні ефективно виконувати свої функції тільки тоді, коли їхня структура достатньо повно відображає форму та зміст одиниць мови, що є об'єктами лексикографування. Тенденція до відтворення цієї повноти насправді спостерігається лише у тих випадках, коли проектування інформаційно-лінгвістичних систем має ґрунтується на глибокому вивченні феноменології мови, яка сама "підказує" вибір адекватного апарату, а також конструкції відповідних моделей. Незважаючи на те, що метою інформаційної науки є витлумачення предметної галузі (у нашому випадку лінгвістичних фактів) мовою моделей даних, самі типи та конструкції цих моделей повинні ви-

²⁹ Соломоник Я. Семантика и лингвистика. — М.: Молодая гвардия, 1995. — 352 с.: ил.

пливати з предметної галузі і якомога точніше враховувати специфіку мовних явищ. Виходячи зі сформульованих положень, було збудовано структурну теорію лексикографічних систем, засновану на феноменології лексикографічного ефекту, послідовне застосування якої надало можливість для необхідних системних узагальнень й створення методології побудови лексикографічних моделей.

Історично початковим пунктом аналізу, результатом якого стало формулювання теорії лексикографічних систем, виявилось дослідження значного числа структур реально існуючих традиційних словників, їхнє узагальнення й побудова відповідних моделей. З метою деталізації та інформаційно-лексикографічної конкретизації було проведено дослідження загальних структуротвірних ефектів та елементів лексикографічних систем, які абстрагуються з традиційних словників, перетворюючись на елементи інфологічних моделей лексикографічних систем “загального положення”. Цей шлях привів до встановлення поняття *структури лексикографічних систем*.

Очевидно, що структура традиційних словників не випадкова, оскільки в ній сконцентровано багатовіковий досвід цілих поколінь лексикографів. Тому вона, як правило, є вільною від суб’єктивних смаків та уподобань розробників інформаційних систем. Досвід лексикографування як різновиду інтелектуальної діяльності (в міру його нагромадження) від систематизації власне філологічних фактів (і навіть ще вужче — від систематизації даних про лексичні одиниці) поступово поширюється й на систематизацію даних про світ, знання про який, у свою чергу, зосереджені у природній мові як цілісній інформаційній системі.

Універсальність явища лексикографічного ефекту спричиняє неодноразово відзначену нами тенденцію до лексикографування будь-якого мовного феномену — саме цей факт пояснює побутування в лексикографічній практиці прикладів створення словників, у яких лексикографуються навіть і такі одиниці мови, які не мають безпосереднього вербального вираження. Так, спроба лексикографування синтаксичних структур зроблена, наприклад, у праці Г. А. Золотової³⁰, у вступі до якої зазначається: “Як фізичний світ навколо нас складається з елементарних частинок, найдрібніших з відомих частинок матерії, так і синтаксична будова нашої мови організується різноманітними, але регулярними комбінаціями елементарних, або мінімальних, одиниць, далі неподільних на синтак-

³⁰ Золотова Г. А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. — М., 1988. — 440 с.

сичному рівні. У лінгвістиці на сучасному етапі її розвитку визріла потреба в осмисленні поняття елементарних синтаксичних одиниць, з яких, як стає все очевидніше, будуються й усі інші, складніші конструкції". І далі: "Синтаксею названо мінімальну, далі неподільну семантико-синтаксичну одиницю російської мови, яка виступає одночасно як носій елементарного смислу і як конструктивний компонент складніших синтаксичних побудов, яка, отже, характеризується певним набором синтаксичних функцій"³¹. Відзначимо очевидну аналогію (подекуди майже текстуальний збіг) з нашим формулюванням лексикографічного ефекту, простір дії якого, очевидно, є незрівнянно ширшим.

Аналогічні спроби лексикографування семантичних структур³² не лише відбивають загальну тенденцію до лексикографічного опису всіх мовних явищ, а й відповідають потребам практики щодо розроблення найдосконаліших систем лінгвістичного забезпечення.

Із викладеного випливає методологічна коректність включення одиниць будь-якого мовного рівня до складу елементарних інформаційно-лексикографічних одиниць тієї чи іншої лексикографічної системи. У такий спосіб лексикографуються семантичні, синтаксичні, когнітивні та інші структури, які, як правило, не мають безпосереднього вербального представлення в системі природної мови. До такого типу робіт близькі й праці зі створення словників: ідеографічних, дієслівного керування, еквівалентів слів, фразеологізмів тощо. До останніх двох типів словників примикає ціла низка можливих лексикографічних праць, поки ще не створених, але які теоретично мають цілковите право на існування³³. У згаданій праці наведено пропозиції до створення понад 50 різних словників, у яких одиницями лексикографування (елементарними інформаційними одиницями щодо відповідних, подекуди вельми екзотичних лексикографічних ефектів) виступають, наприклад: звернення, етикетні фрази, гоноративи (висловлювання ввічливості), гумілятиви (вирази хамства), стимули і реакції (підтакування, згоди, заперечення, спростування) тощо.

Вивчення різноманітних структур існуючих традиційних словників дозволяє зробити певні узагальнення, які можна не лише по-

³¹ Золотова Г. А. Знач. праця. — С. 3—4.

³² Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Под общ. ред. Н. Ю. Шведовой. — М., 1998. — Т. I. — 832 с.

³³ Девкин В. Д. О неродившихся немецких и русских словарях // Вопросы языкознания. — 2001. — № 1. — С. 85—97.

класти в основу теоретичної лексикографічної схеми, а й використати при проектуванні конкретних інформаційних систем лінгвістичного напрямку й при створенні відповідного програмного забезпечення. Оскільки в лексикографії вже давно відбулося розмежування понять “словник” та “список слів”, “перелік”, “індекс”, “інвентар”, словник як абстрактна лексикографічна система обов’язково наділений структурою, що містить мінімум дві необхідні частини: реєстрову (ліву) та інтерпретаційну (праву), що є виявом ВФЗ. Саме наявність інтерпретаційної частини — носія змістового компонента ВФЗ — відрізняє словник від звичайного списку слів. Але словник має і глибшу структуру, яка відбивається в будові реєстрової та інтерпретаційної частин словника в цілому та його окремих словникових статей, а також у структурі міжстатейних та міжсловникових відображень. Відтак словник являє собою спеціальний вид тексту, в якому в систематизованому та структурованому вигляді подається опис лексики певної мови (або сукупності мов). Однак словник природно розглядати і як специфічний об’єкт техніки, а саме — інформаційну систему, де через поліграфічне виконання означено певні лінгвістичні ефекти за допомогою шрифтових виділень, позиційного розміщення, спеціальних позначок тощо, які відіграють роль ідентифікаторів відповідних інформаційних змінних — елементів метамови словника. Складність будови словника полягає ще й у тому, що не всі елементи його структури явно означені вказаним вище способом. У структурі реальних словників, як правило, існує велика кількість неявних структуротвірних елементів, виявлення яких часто є досить складним завданням. Процес абстрагування словникової (лексикографічної) структури становить своєрідне розшифрування, реконструкцію того лексикографічного ефекту, який спричинив утворення цієї структури, і розгортається з використанням кількох положень, сформульованих спершу в мовознавстві, але по суті таких, що мають загальносистемний характер.

Побудова структурної моделі лексикографічних (словникових) систем орієнтується на *багатоаспектне представлення* знакової природи лексичних одиниць як найкомпактніших і найінформативніших у природній мові. З позицій теорії лексикографічного ефекту це означає виділення в досліджуваній мовній системі підсистеми її елементарних інформаційних одиниць і визначення множини їх системно-структурних параметрів.

Наступний момент полягає в урахуванні *дихотомічної структури* кожної елементарної інформаційної одиниці (і повної їх су-

купності), що відбивається в багатовимірному співвідношенні форми та змісту, носієм якого є визначений клас елементарних інформаційних одиниць.

Багатоаспектність представлення знакової природи одиниць природної мови в традиційних словниках (або елементарних інформаційних одиниць в загальних лексикографічних системах) забезпечується врахуванням семіологічних, лінгвістичних (фонетичних, морфематичних, граматичних, семантичних, стилістичних та ін.) і когнітивних особливостей об'єктів лексикографування — залежно від типу словника й глибини характеристизації лексикографічного ефекту, який виступає предметом дослідження в кожному конкретному випадку. В інформаційно-лексикографічній моделі з зазначеними особливостями зіставляється певна множина комплексів даних та/або знань.

Зауважимо, що в мовному (мовленневому) потоці онтологічна природа мови виступає неподільною на окремі складові частини, які є в концептуальних представленнях. З цього випливає прагнення до створення "інтегральних" словників і, отже, необхідність використання комплексних ("інтегрованих") моделей мовних явищ, тому при розробленні комп'ютерних систем опрацювання мови постає завдання створення таких формалізованих моделей, які були б налаштовані на ефективне представлення інтеграційних процесів і водночас урахували б специфіку лінгвістичних об'єктів. Таким чином, критерій багатоаспектності репрезентації знакової природи одиниць мови дозволяє побудувати комплексні, інтегровані моделі даних, придатні до поєднання концептуальних представлень різних за своєю природою мовних явищ.

Дихотомічність структури елементарних інформаційних одиниць в інформаційно-лексикографічній моделі (подібно до того, як це робиться в більшості традиційних словників) виявляється в структурній організації лексикографічної системи й впливає з фундаментальних положень сучасної лінгвістики, яка оперує поняттями форми і змісту, внутрішньої та зовнішньої форми мовних одиниць, феноменологія яких глибоко простежена на мовному матеріалі.

Як відзначає В. М. Русанівський³⁴, мова має дуалістичну функцію: з одного боку, це матеріальна основа, на яку мислення спирається в процесі свого функціонування, а з другого — матеріал, у якому воно фіксується, стаючи dokonаним фактом. Об'єктами до-

³⁴ Русанівський В. М. Структура лексичної та граматичної семантики. — К., 1988. — 240 с.

слідження складників мовнорозумового потоку є як фізична ("матеріальна"), так і змістова ("ідеальна") сторони. Отже, звукову субстанцію мовлення можна вважати його формою, а інформаційні властивості — змістом. З огляду на цю обставину, звукова реалізація мовлення може поділятися на елементи різного ступеня агрегованості — інтонаційно цілісні одиниці (інтонеми), комбінації голо- сних і приголосних (склади), власне голосні та приголосні (звуки) і т. д. Цей процес необмежений, оскільки виділення та класифікація складників звуків мовлення залежать від цілого ряду чинників, у тому числі й від прогресу акустики, фонології тощо. Фізичний про- цес мовлення належить до незворотних і (як і багато інших акусти- чних явищ) дисипативних процесів. Зазначені властивості фізичної субстанції мовлення разом з властивостями мовленнєвого апарату визначають її зовнішні інформаційні характеристики.

У свою чергу, писемна форма мови моделює її усну форму, то- му взагалі справедлива послідовність: <модель дійсності — мис- лення> → <модель мислення — усна мова> → <модель усної мови — писемна мова>. Оскільки наведені моделі фізично реалізуються в єдиній системі (пов'язаній з індивідуумами, соціальними спільно- тами, системами культури тощо), цілком природні та необхідні їх взаємодія та взаємовпливи. Таким чином, писемний варіант мови так само виступає у функції і моделі мислення, і моделі дійсності.

Безпосереднє буття мови у вигляді мовленнєвої діяльності, а та- кож писемності та інших способів фіксації мовних актів на фізич- них носіях, відмінних від природномовних, репрезентує власти- вість мови "мати зовнішню форму". Зовнішня форма взагалі мож- лива завдяки здатності мови бути "представником" феноменальної сторони дійсності, і оскільки мова сама становить певну дійсність, у ній закладені потенції для позначення ("представлення") самої себе.

Система, яка виступає репрезентантом феноменальної сторони дійсності, повинна бути й певним чином організованою. Оскільки відмінність між явищем і сутністю є відносною, а між феноменаль- ною і сутнісною дійсністю немає чіткої межі, її не повинно бути і в мові як моделі дійсності. Цей факт і реалізується у властивості сло- ва мати внутрішню форму, яка пов'язана з представленням ноуме- нальної частини його власного буття саме в системі мови. Зовнішня і внутрішня форми, таким чином, взаємопов'язані й разом станов- лять форму слова на противагу його змісту — сумі конкретних значень.

Усе це дає підстави для твердження, що ВФЗ (включаючи й уяв- лення про внутрішню та зовнішню форму мовних одиниць) мають

загальний характер, репрезентують універсальну властивість елементарних інформаційних одиниць, що індукуються в процесі розвитку будь-якого лексикографічного ефекту, і, будучи формалізованими у вигляді відповідних моделей даних, спроможні утворити субстрат моделей інформаційних систем довільної природи та походження, а для мовно орієнтованих моделей даних вони взагалі є обов'язковими. Зазначені поняття, на наш погляд, мають потенціал конструктивності, оскільки будь-який зміст існує лише у певній формальній оболонці, що дозволяє застосовувати уніфікований підхід до побудови їх репрезентантів у науковій теорії.

Розглянемо певну систему D , концептуальний опис якої представимо у вигляді певної лексикографічної системи. Оскільки нас цікавлять насамперед лінгвістичні факти, системою D у нас виступатиме саме природна мова або сукупність природних мов, або певна підсистема (певний аспект) природної мови.

Відповідно до викладеного системі D властива складна ієрархія лексикографічних ефектів. Так, для природномовної системи можна навести приклади цілої низки лексикографічних ефектів, результатом яких постало виділення з мовленнєвого потоку окремих фонем, складів, морфем, слів (словоформ), лексем, словосполучень, еквівалентів слів, речень тощо. Усі названі структурні одиниці виступають як складники відповідних класів елементарних інформаційних одиниць стосовно тих чи інших типів природномовних лексикографічних ефектів.

У подальшому як *лексикографічну систему (Л-систему)* розглядатимемо спеціальне інформаційне середовище, в якому розвивається (реалізується) певний лексикографічний ефект (або певна сукупність лексикографічних ефектів).

Для побудови практично корисної схеми моделювання наведених явищ необхідно визначити набір інформаційних конструктивів, які специфікують структурні елементи Л-систем, дозволяючи розробляти конкретні застосування. У свою чергу, це спричиняє необхідність побудови конструктивної теорії Л-систем — в її основу покладено лексикографічну модель даних, розроблену в наших працях³⁵, результатами і позначеннями яких ми скористаємося у подальшому викладі.

³⁵ Широков В.А. Інформаційна теорія лексикографічних систем; Широков В.А., Рабулець О.Г. Формалізація у галузі лінгвістики // Актуальні проблеми української лінгвістики: теорія і практика. — К., 2002. — Вип. 5. — С. 3-27.

Відповідно до інформаційної інтерпретації процесів сприйняття³⁶ визначимо результат рецепції суб'єктом S класу елементарних інформаційних одиниць (ЕІО) $I^Q(D)$ у вигляді певної множини $V(I^Q(D))$ — множини описів одиниць, що належать до класу $I^Q(D)$; ця множина є результатом процесу:

$$S : I^Q(D) \rightarrow V(I^Q(D)), \quad (1.12)$$

тому для кожного елемента $x \in I^Q(D)$ однозначно визначено його опис $V(x)$ як елемент множини $V(I^Q(D))$: $V(x) \in V(I^Q(D))$; $Sx = V(x)$. Отже, логічно припустити, що $V(I^Q(D))$ має вигляд об'єднання:

$$V(I^Q(D)) = \bigcup_{x \in I^Q(D)} V(x). \quad (1.13)$$

Згідно з інформаційною концепцією представлення опису системи ЕІО, кожний $V(x)$ зображується у вигляді слова (тексту) в певному скінченному алфавіті $A = \{a_1, a_2, \dots, a_n\}$, тобто скінченної послідовності символів з A . Надалі слова в алфавіті A називатимемо A -словами. Наприклад, якщо ми розглядаємо Словник української мови³⁷ то його алфавіт A складається з таких елементів:

- звичайний алфавіт української мови (великі і малі літери);
- знаки пунктуації;
- арабські цифри;
- римські цифри;
- символи пробілу та абзацу;
- спецсимволи ($//$, Δ , \blacktriangle , \diamond , \blacklozenge ,...);
- типи шрифтів тощо.

У такий спосіб опис $V(x)$ будь-якої елементарної інформаційної одиниці x , $x \in I^Q(D)$, представляється A -словом такого вигляду:

$$V(x) = v_1(x)v_2(x)\dots v_{k(x)}(x), v_i(x) \in A, i = 1, 2, \dots, k(x), k(x) \geq 1. \quad (1.14)$$

де кожна "літера" $v_i(x)$ (A -літера) береться з алфавіту A . Відзначимо, що довжина $k(x)$ A -слова $V(x)$ залежить від x — тобто від того, яким обрано елемент $x \in I^Q(D)$. Формула (1.14), за визначенням, подає достатньо повний, у певному смислі вичерпний опис елементарної інформаційної одиниці x в даній лексикографічній системі. Між класом ЕІО $I^Q(D)$ і множиною описів $V(I^Q(D))$ за допомогою

³⁶ Широков В.А., Рабулець О.Г. Знач. праця.

³⁷ Словник української мови в 11 томах. — К.: Наукова думка. — 1970—1980.

відображення δ встановлюється певний ізоморфізм. Іншими словами, множина описів $V(I^Q(D))$ є певною власною підмножиною множини $W(A)$: $V(I^Q(D)) \subset W(A)$, причому $W(A)$ являє собою множину всіх слів скінченної довжини над A , тобто послідовностей вигляду $v_1 v_2 \dots v_q$, $q < \infty$, $v_i \in A$, $i = 1, 2, \dots, q$. Вважаємо, що слово нульової довжини — 0 також належить до $W(A)$: $\forall a \in W(A) \exists 0 \in W(A)$, таке, що $a * 0 = 0 * a = a$, де “*” — конкатенація. Замкненість відносно операції конкатенації, тобто вимога: $a, b \in W(A) \Rightarrow \exists c \in W(A)$, $c = a * b$, а також асоціативність відносно неї $\forall a, b, c \in W(A) \Rightarrow a * b * c = (a * b) * c = a * (b * c)$ перетворює $W(A)$ на напівгрупу з напівгруповою операцією “*” і одиничним елементом 0 .

Вибір алфавіту A , через який зображено $W(A)$ та $V(I^Q(D))$, тут не обґрунтовуватимемо й не конкретизуватимемо, що відповідає алгебраїчній традиції, проте зауважимо, що його генерація є наслідком певного лексикографічного ефекту, який розвивається в системі мовлення (акустичного) та його інформаційно-графічної інтерпретації. Якщо ми розглядаємо звичайні словники, то природною є інтерпретація A -слова $V(x)$ як тексту словникової статті із реєстровою одиницею x .

Взагалі структура напівгрупи є досить бідною, а конструкція $W(A)$ — занадто широкою для ефективного виявлення в ній характерних властивостей мовних систем. Для досягнення цієї мети необхідно запровадження додаткових припущень та обмежень, за допомогою яких у структурі $W(A)$ виділяються підструктури, характерні саме для природної мови. Це досягається в такий спосіб.

Оскільки кожний $V(x)$, як зазначалося, є адекватним та однозначним репрезентантом (описом) відповідного елемента x із системи $I^Q(D)$, в його структурі повинні бути з достатньою повнотою відображені властивості цього елемента. Враховуючи лінійний характер об'єкта $V(x)$, котрий зображається лінійною послідовністю символів з A , доходимо висновку, що єдиним природним джерелом його структури може виступати лише певна множина його A -підслів та певні відношення між її елементами. A -підслова в описі $V(x)$ визначаються як A -слова, складені з тих символів алфавіту A , що містяться в розглядуваному описі $V(x)$ і розташовуються в A -підслові у порядку, індукованому порядком розташування літер у самому описі. Очевидно, що множина всіх A -підслів A -слова довжини n (тобто A -слова, що складається з n A -літер) містить 2^n елементів. Позначимо множину всіх A -підслів A -слова $V(x)$ через $B[V(x)]$.

Структура на множині описів вводиться у такий спосіб. Припустимо, що для всіх описів $V(x)$ існує єдине правило, за яким з будь-якого A -слова $V(x)$ можна виділити множину A -підслів $\beta(x) = \{\beta_i(x)\}$ з такими властивостями:

- елемент x належить до множини $\beta(x)$;
- весь опис $V(x)$ є елементом множини $\beta(x)$;
- правило, за яким виділяються елементи множини $\beta(x)$ є єдиним для всіх $V(x)$.

Описаним способом з будь-якого $V(x)$ виділяється множина $\beta[V(x)]$ величин (A -підслів) $\beta_i(x)$ такого вигляду:

$$\beta[V(x)] \equiv \{\beta_i(x), i = 1, 2, \dots, q\} \subseteq B[V(x)], \quad (1.15)$$

де $B[V(x)] = \{v_{i_1} v_{i_2} \dots v_{i_p}, 1 \leq i_1 < i_2 < \dots < i_p \leq k(x), p = 1, 2, \dots, k(x)\}$, причому:

$$v_{ij} \in \{v_{1(x)}, v_{2(x)}, \dots, v_{k(x)}(x)\}; x \in \beta[V(x)]; V(x) \in \beta[V(x)],$$

$$\beta_{k(x)} \neq \beta_{m(x)} \text{ при } k \neq m. \quad (1.16)$$

Покладемо за визначенням:

$$\beta[V(I^Q(D))] = \bigcup_{x \in I^Q(D)} \beta[V(x)]. \quad (1.17)$$

Очевидно, що $V(I^Q(D)) \in \beta[V(I^Q(D))]$. Позначимо

$$\beta_i = \bigcup_{x \in I^Q(D)} \beta_i(x), i = 1, 2, \dots, q, \text{ а також } \beta = \bigcup_i \beta_i. \quad (1.18)$$

Зрозуміло, що $\beta \equiv \beta[V(I^Q(D))]$. Відзначимо, що деякі з елементів $\beta_i(x), i = 1, 2, \dots, q$, можуть бути порожніми для певних $x \in I^Q(D)$; у цьому випадку вони випускаються у формулах (1.15)—(1.18).

Через $\sigma[\beta]$ позначимо певну структуру, визначену на β і, отже, на $V(I^Q(D))$,— надалі називатимемо $\sigma[\beta]$ макроструктурою $V(I^Q(D))$; обмеження $\sigma[\beta]$ на $V(x)$: $\sigma[\beta] \upharpoonright_{V(x)} \equiv \sigma(x)$ породжує мікроструктуру $V(x)$. Активне формулювання цього факту полягає у встановленні процедури (оператора, процесу...) σ , який породжує на β структуру $\sigma[\beta]$:

$$\sigma: \beta \rightarrow \sigma[\beta]. \quad (1.19)$$

На β можлива генерація цілої низки неізоморфних структур $\sigma[\beta]$. У їх ролі можуть виступати будь-які з відомих моделей даних (ієрархічна, мережева, реляційна, об'єктно-реляційна та ін.), логіко-

математичні моделі (зокрема, логічні числення типу логіки предикатів), конструкції формальних граматики тощо. Одним із можливих механізмів формування структури може бути такий. Збудуємо таблицю:

β_1	β_2	...	β_q
$\beta_1(x_1)$	$\beta_2(x_1)$...	$\beta_q(x_1)$
$\beta_1(x_2)$	$\beta_2(x_2)$...	$\beta_q(x_2)$
...
$\beta_1(x_M)$	$\beta_2(x_M)$...	$\beta_q(x_M)$

Деякі з елементів $\beta_i(x_j)$, очевидно, можуть бути порожніми, тому довжини стовпчиків таблиці, взагалі кажучи, є різними. Величини β_i , $i = 1, 2, \dots, q$, інтерпретуватимемо як атрибути (імена атрибутів), а набори $\text{Dom } \beta_i = \{\beta_i(x_1), \beta_i(x_2), \dots, \beta_i(x_M)\}$, $i = 1, 2, \dots, q$, як домени цих атрибутів. Тоді структура $\sigma[\beta]$ може бути реалізована у вигляді певної реляційної алгебри $R[\beta]$, визначеної на декартовому добутку:

$$\bigtimes_{i=1}^q \text{Dom } \beta_i = \beta^{\otimes} \quad (1.20)$$

Іншими словами, якщо структура σ ототожнюється з певною реляційною алгеброю R над β^{\otimes} , то трійка $\{V(I^{\otimes}(S)), \beta, R[\beta]\}$ представляє собою ніщо інше, як реляційну модель, а п'ятірка $\{I^{\otimes}(S), D, V(I^{\otimes}(S)), \beta, R[\beta]\}$ — задає деяку об'єктно-реляційну модель³⁸. При цьому $I^{\otimes}(S)$ представляє клас об'єктів моделі; β_i інтерпретуються як атрибути (імена атрибутів) з доменами $\text{Dom } \beta_i$, елементами котрих служать $\beta_i(x)$, $x \in I^{\otimes}(S)$. Зрозуміло, що окремими доменами є сама множина $\{x\}$ (для скорочення викладу ми без подальшої деталізації ототожнюємо елемент x як такий, що належить до класу $I^{\otimes}(S)$, з його "ім'ям" в $V(x)$), і $V(I^{\otimes}(S))$ — як множину $\{V(x)\}$. Реляційні відношення відповідних арностей визначаються як завжди — у вигляді певних підмножин множини β^{\otimes} . Їх кортежами виступають елементи вигляду:

³⁸ Коннолли Томас, Бегг Каролін, Страчан Анна. Базы данных: проектирование, реализация и сопровождение. Теория и практика. — 2-е изд.: Пер. с англ. — М.: Издательский дом "Вильямс", 2000. — 1120 с.: ил.

$$(\beta_{i1}(x_{j1}), \beta_{i2}(x_{j2}), \dots, \beta_{ir}(x_{jir})), i_1 < i_2 < \dots < i_r; x_{jim} \in I^Q(S), m=1, 2, \dots, r. \quad (1.21)$$

Реляційне числення в цій моделі визначається звичайним чином (див., наприклад³⁹).

Завдяки можливості інтерпретації $I^Q(D)$ як класів, елементами яких виступають об'єкти будь-якого походження, природною є об'єктно-орієнтована інтерпретація моделі. Відношення між елементами класу $I^Q(D)$ індукуються системою унарних відношень $r[\beta_i]$ на $\beta_i = \{x\}$ та відображенням:

$$\Delta: V(I^Q(D)) \rightarrow I^Q(D); \Delta r[\beta_i]. \quad (1.22)$$

У такий спосіб комплекс $I^Q(D)$ постає представником онтологічної природи дійсності, яка підлягає моделюванню, у той час як $V(I^Q(D))$, β , $\sigma[\beta]$ репрезентують її концептуальну сторону.

Подальший розгляд концентруватиметься навколо ВФЗ, носієм яких є комплекс $I^Q(D)$ і які розвиваються й реалізуються в середовищі $\{I^Q(D), S, V(I^Q(D)), \beta, \sigma[\beta]\}$. При цьому сукупність властивостей і якостей комплексу $I^Q(D)$, згідно з викладеним вище, розподіляється на дві не цілком чіткі й не дуже виразно відокремлені частини. Зауважимо, що в концептуальній схемі, реалізованій в описі $V(I^Q(D))$, вказані частини повинні бути відокремлені; інакше кажучи, необхідною умовою коректності її побудови є існування процедури, яка здійснює таке відокремлення. Усе це може бути відображене в комутативній діаграмі:

$$\begin{array}{ccc}
 & & V(I^Q(D)) \\
 & \swarrow & \searrow \\
 F & \xrightarrow{H} & C \\
 \Lambda(I^Q(D)) & & P(I^Q(D))
 \end{array} \quad (1.23)$$

$FV(I^Q(D)) = \Lambda(I^Q(D)); CV(I^Q(D)) = P(I^Q(D)); \Lambda(I^Q(D)) \cap P(I^Q(D)) = \emptyset$, причому $H \circ F = C$, де символом "o" позначено композицію відображень;

$$\Lambda(I^Q(D)) = \bigcup_{x \in I^Q(D)} \Lambda(x); P(I^Q(D)) = \bigcup_{x \in I^Q(D)} P(x). \quad (1.24)$$

На $\Lambda(I^Q(D))$ і $P(I^Q(D))$ індукуються макроструктури:

³⁹ Ульман Дж. Основы систем Баз данных. М.: Финансы и статистика, 1983. — 334 с.

$$F\sigma[\beta] = \lambda[\beta] \text{ і } C\sigma[\beta] = \rho[\beta] \quad (1.25)$$

і відповідні мікροструктури:

$$\lambda[\beta] \big|_{V(x)} \equiv \lambda(x); \rho[\beta] \big|_{V(x)} \equiv \rho(x) \quad (1.26)$$

як обмеження $\lambda[\beta]$ і $\rho[\beta]$ на $V(x)$.

Відзначимо, що діаграма (1.23), тобто об'єкти $V(I^Q(D))$, $\Lambda(I^Q(D))$, $P(I^Q(D))$ і відображення F , C , H , фактично будується незалежно від структури $\sigma[\beta]$. Походження і зміст її складових елементів зовсім інші. А саме: $\Lambda(I^Q(D))$ співвідноситься з тією частиною опису $V(I^Q(D))$, яка — в певному розумінні — представляє форму $I^Q(D)$, а $P(I^Q(D))$, відповідно, зставляється з тією частиною опису $V(I^Q(D))$, що відповідає за зміст $I^Q(D)$. Наведена специфікація практично підтверджує думку про те, що ВФЗ є універсальними, притаманними об'єктам будь-якого походження, оскільки вони залучаються до процесу взаємодії з суб'єктом, який їх досліджує.

Визначення 1. Вісімка об'єктів $\{I^Q(D), S, V(I^Q(D)), \beta, \sigma[\beta], F, C, H\}$ визначає елементарну лексикографічну модель даних, а її конкретна реалізація — елементарну лексикографічну систему. Інколи для скорочення, коли не виникатиме різночитань, будемо позначати через $V(I^Q(D))$ цілу елементарну лексикографічну систему.

Зауважимо, що будь-який елемент β_i (або будь-яка їх сукупність), який належить до структур β , $\sigma[\beta]$, $\lambda[\beta]$, $\rho[\beta]$, може бути інтерпретований як елементарна Л-система. Звідси одержуємо можливість виокремити в структурі початкової елементарної Л-системи низку інформаційно-лінгвістичних підструктур, які тлумачимо як окремі Л-системи. Перевизначивши у такий спосіб структуру вихідної Л-системи, одержуємо Л-модель та Л-систему загального положення (не елементарну); вона представлена у вигляді об'єднання певної кількості елементарних Л-систем з можливими відображеннями та зв'язками між ними. Таким чином, Л-система загального положення має вигляд графа $G = \{V = \{V_i\}; R = \{R_{kl}\}\}$, де $V = \{V_i\}$ — множина вершин, якими виступають елементарні Л-системи V_i , що входять до G , а $R = \{R_{kl}\}$ — множина ребер графа G , R_{kl} поєднує V_k та V_l .

Зокрема, ніщо не заважає розглядати $\Lambda(I^Q(D))$ і $P(I^Q(D))$ як окремі, автономні елементарні Л-системи, а це уможливило таку побудову:

$$\begin{array}{ccc}
 V(I^Q(D)) = (\Lambda(I^Q(D)) \equiv \Lambda_0(I^Q(D))) & \xrightarrow{H_0} & P_0(I^Q(D)) \equiv P(I^Q(D)) \\
 \begin{array}{cc}
 \swarrow & \searrow \\
 F_{01}^\Lambda & C_{01}^\Lambda
 \end{array} & & \begin{array}{cc}
 \swarrow & \searrow \\
 F_{01}^P & C_{01}^P
 \end{array} \\
 \Lambda_{01}^\Lambda(I^{Q1}(D)) \xrightarrow{H_{01}^\Lambda} P_{01}^\Lambda(I^{Q1}(D)) & & \Lambda_{01}^P(I^{Q2}(D)) \xrightarrow{H_{01}^P} P_{01}^P(I^{Q2}(D))
 \end{array} \quad (1.27)$$

Звернімо увагу на зміну типу лексикографічного ефекту на другому поверсі — замість Q тепер маємо Q_1 й Q_2 відповідно. Отже, приходимо до комплексів об'єктів $I^{Q1}(D)$ та $I^{Q2}(D)$. Продовжуючи цей процес, одержуємо рекурсивне розвинення лексикографічної системи $V(I^Q(D))$:

$$\begin{array}{ccc}
 & V = (\Lambda_0; P_0) & \\
 & \swarrow & \searrow \\
 \Lambda_0 & & P_0 \\
 \swarrow & & \searrow \\
 \Lambda_{01}^\Lambda & & P_{01}^P \\
 \swarrow & & \searrow \\
 \dots & & \dots
 \end{array} \quad (1.28)$$

Назвемо цей процес *рекурсивною редукцією лексикографічної системи*. Він нагадує своєрідний інформаційний “мікроскоп”, що виявляє все тонші деталі структури лексикографічної системи.

Надалі позначатимемо процес рекурсивної редукції Л-системи $V(I^Q(D))$ через $RR \downarrow [V(I^Q(D))]$. У визначення цього процесу входять характеристики всіх операторів F, C, H на всіх наявних рівнях рекурсивної редукції разом із результатами їх дії, а також усі макро- і мікруктури σ, λ, ρ .

Викладена конструкція становить зміст загального визначення лексикографічної моделі даних:

$$\{I^Q(D), S, V(I^Q(D)), \beta, \sigma [\beta], RR \downarrow [V(I^Q(D))]\} \quad (1.29)$$

та лексикографічної системи:

$$\{I^Q(D), S, V(I^Q(D)), \beta, \sigma [\beta], RR \downarrow [V(I^Q(D))], \Sigma\}, \quad (1.30)$$

де символом Σ позначено її архітектуру як інформаційної системи.

Визначення окремих елементів моделі у формулах (1.29)—(1.30) представлено формулами (1.23)—(1.28). Архітектура Σ , зазвичай, обирається трірівневою, такою що відповідає ANSI/X3/SPARK (або просто ANSI/SPARK)⁴⁰. У дещо іншому формулюванні аналогічну архітектуру було запропоновано у звіті⁴¹, де також запроваджено три рівні абстракції даних: “концептуальний” — “фізичний” — “представлен”. З певними уточненнями можна вважати справедливою відповідність названих трьох рівнів концептуальному, внутрішньому та зовнішньому рівням архітектури ANSI/SPARK. Слід згадати також працю Цикритзиса і Клуґа⁴², яка є неформальним вступом до переглянутої версії звіту⁴⁰. Три рівні абстракції представлені у великій кількості існуючих баз даних. Основні складові архітектури ANSI/SPARK використовуватимемо в такій інтерпретації:

$$ARCH_LS = \{CM, EXM, INM; \Phi, \Psi, \Xi\}, \quad (1.31)$$

де символом CM позначено концептуальну модель лексикографічної системи LS . Символом $EXM = \{exM\}$ позначено множину її зовнішніх моделей, які відповідають даній концептуальній моделі CM , а $INM = \{inM\}$ — відповідна множина її внутрішніх моделей. Через $\Phi = \{\phi\}$ позначено множину відображень CM в EXM :

$$\phi: CM \rightarrow exM, \text{ де } exM \in EXM; \quad (1.32)$$

відповідно $\Psi = \{\psi\}$ — множина відображень CM в INM :

$$\psi: CM \rightarrow inM, \text{ де } inM \in INM; \quad (1.33)$$

$\Xi = \{\zeta\}$ — множина відображень INM в EXM :

$$\zeta(inM) = exM. \quad (1.34)$$

При цьому ми зупиняємося на такій інтерпретації елементів архітектури.

Концептуальна модель має такі властивості:

1. *Семіотичність*. Концептуальна модель реалізована в середовищі певної знакової системи. Її побудова ґрунтується на спільному розгляді предметної галузі, галузі мислення і знакової галузі.

⁴⁰ ANSI/X3/SPARK DBMS study group interim report. // FDT-Bull. ACM SIGMOD. — 1975. — V. 7. — №2. — 140 p.

⁴¹ CODASYL DBTG 1971 [CODASYL Data Base Task Group April 71 Report. ACM New York, 1971]

⁴² Tsichritzis D. and Klug A. (eds.) The ANSI/X3/SPARK Framework, AFIPS Press, Nontvale, N. J., 1978.

2. *Семантичність*. В категоріях моделі відображаються об'єкти та зв'язки між ними, суттєві з точки зору адекватного опису знань про предметну галузь. Модель не повинна залежати від логічної, фізичної та зовнішньої репрезентації даних.

3. *Однозначність*. Концептуальна модель описує предметну галузь згідно з принципом однозначного іменування, згідно з яким кожен знак моделі має одне значення, один смисл. Омонімія і полісемія у визначенні елементів моделі є знятою. Використання кваліфікацій і контекстних механізмів не допускається.

4. *Несуперечливість*. Для кожного стану концептуальної моделі межі між її категоріями є абсолютними: при класифікації чи описі об'єктів предметної галузі кожному з них відповідає однозначно визначена множина станів, які попарно не перетинаються.

5. *Інтегрованість*. У концептуальній моделі суміщаються уявлення різних фахівців про предметну галузь. Суперечливість цих уявлень фіксується за допомогою засобів обмеження цілісності і усувається через посередництво спеціальних процедур.

6. *Типізованість*. Всі концептуально визначені властивості об'єктів моделі повинні мати інтерпретацію на певних типах даних.

7. *Алгоритмізованість*. Всі об'єкти концептуальної моделі, а також зв'язки, відображення та операції над ними повинні мати скінченний опис, тобто мати інтерпретацію у вигляді алгоритмів скінченної складності над конструктивними об'єктами.

Отже, концептуальна модель (*концептуальний рівень представлення*) предметної галузі — це знакова, семантична модель, у якій в однозначному, скінченному та несуперечливому вигляді інтегруються уявлення різних фахівців про предметну галузь.

У внутрішній моделі (*внутрішньому рівні представлення*) визначаються типи, структури та формати представлення, зберігання та маніпулювання даними, алгоритмічна база та операційно-програмне середовище, в яке "занурюється" концептуальна модель при її реалізації.

Зовнішня модель (*зовнішній рівень представлення*) відображає погляди кінцевих користувачів і, отже, прикладних програмістів на інформаційну систему. В ній реалізується система засобів, які дозволяють користувачеві здійснювати дозволені контакти і маніпулювання даними, представленими у внутрішньому рівні.

Одній концептуальній моделі може відповідати декілька внутрішніх та зовнішніх моделей, тому відображення $\Phi: SM \rightarrow EXM$ та $\Psi: SM \rightarrow INM$, взагалі кажучи, не ін'єктивні. Але відображення $\phi: SM \rightarrow exM$, та $\psi: SM \rightarrow in M$ будуються ін'єктивними (але,

зазвичай, не бієктивними). Визначимо множину Ξ відображень INM в EXM :

для $\forall inM \in INM$ та $\forall exM \in EXM \exists \xi \in \Xi$ таке, що:

$$\xi(inM) = exM. \quad (1.35)$$

При цьому відображення φ, ψ, ξ будуються в такий спосіб так, що діаграма:

$$\begin{array}{ccc}
 CM & \xrightarrow{\psi} & inM \\
 & \searrow \varphi & \downarrow \xi \\
 & & exM
 \end{array} \quad (1.36)$$

є комутативною: $\xi \circ \psi = \varphi$. Вимога комутативності цієї діаграми є суттєвою, оскільки гарантує узгодженість між усіма рівнями архітектури системи. У свою чергу:

$$CM = \{Ob(LS); RelOb(LS); Mor(Ob(LS), RelOb(LS));$$

$$Res(Ob(LS), RelOb(LS); Mor(Ob(LS), RelOb(LS)))\}, \text{ де:}$$

$Ob(LS)$ — множина об'єктів та категорій LS ; $RelOb(LS)$ — множина зв'язків (відношень) між об'єктами та категоріями LS ; $Mor(Ob(LS), RelOb(LS))$ — множина можливих операцій (процесів) над $Ob(LS)$ та $RelOb(LS)$; $Res(Ob, RelOb, Mor(Ob, RelOb))(LS)$ — множина обмежень цілісності. Всі названі елементи мають однозначну інтерпретацію в термінах лексикографічної моделі даних $\{I^{\rho}(S), d, V(I^{\rho}(S)), \beta, \sigma[\beta], RR \downarrow \{V(I^{\rho}(S))\}\}$.

Внутрішня модель лексикографічної системи містить такі елементи:

$$INM = \{D(t, s, f), ALG(D(t, s, f)), OS, PL\}, \text{ де:}$$

$D(t, s, f)$ — множина даних, специфікована за типами, структурами та форматами t, s, f , відповідно, за допомогою яких у внутрішньому рівні представляються елементи CM ; $ALG(D(t, s, f))$ — множина алгоритмів (процесів) обробки та маніпулювання даними; OS — множина операційних платформ та PL — множина мов програмування (включаючи і процедурні мови типу СКБД), на яких реалізовано $D(t, s, f)$ та $ALG(D(t, s, f))$.

Зовнішня модель лексикографічної системи представляється у вигляді:

$$EXM = \{IF, SC, FUNC, PROC, APR\}, \text{ де:}$$

IF — інтерфейс системи; *SC* — множина сценаріїв; *FUNC* — множина функцій; *PROC* — множина допустимих процесів, *APR* — множина прикладних програм.

Конкретну інформаційно-лінгвістичну реалізацію лексикографічної системи у певному комп'ютерному середовищі надалі називатимемо лексикографічною базою даних (ЛБД).

1.4.3. Лексикографічні структури і словники

У застосуванні до традиційних словників та словникових комплексів лексикографічні системи набувають простої та прозорої інтерпретації. Справді, алфавіт Л-системи ототожнюється зі знаковою системою словника (включаючи й спецсимволи), клас ЕЮ $I^{\rho}(D)$ — з множиною реєстрових одиниць (об'єктів лексикографування — $\{x\}$), множина описів $V(I^{\rho}(D)) = \{V(x)\}$ — з множиною словникових статей, де заголовкові одиниці пробігають множину $\{x\}$, а $\Lambda(x)$ та $P(x)$ — з лівими та правими частинами відповідних словникових статей, і т.д.

Приклади побудови лексикографічних систем до конкретних лексикографічних праць буде наведено у наступних розділах. Ми переконуємося, що будь-який традиційний словник можна представити у вигляді певної Л-системи. Водночас, застосування конструктивів Л-систем надає засоби для далеко йдучих узагальнень традиційних проблем лексикографії, а багато з нерозв'язаних її проблем набувають свого природного розв'язку у парадигмі Л-систем.

Наведемо для прикладу проблему упорядкування словникових статей за різними критеріями, яка у традиційному словнику не має природного розв'язку. У Л-системі зазначена проблема зводиться до задання системи класифікацій на множині $I^{\rho}(D)$, яка слугить основою для побудови відповідного пошукового апарату і реалізується відповідними засобами внутрішньої та зовнішньої моделі Л-системи. Найпростіша з таких класифікацій — абеткова — породжує так зване лексикографічне упорядкування множини $I^{\rho}(D)$ і, отже, всього словника. Морфемна класифікація, яка полягає у виділенні класів слів з однаковими основами, веде до гніздового принципу, який є характерним для словників словотвірного типу, а також використовується у двомовних перекладних словниках. Граматична класифікація може породити цілу гаму словників. Цей перелік можна продовжити. Зрозуміло, що традиційний, паперовий

словник природно упорядковується лише за одним із класифікаційних принципів (хоча існують і спроби поєднання різних принципів класифікації в одному словникові). Зовсім інші можливості відкриваються для загальних лексикографічних систем та їх реалізацій — комп'ютерних словників, пошуковий апарат яких може одночасно включати всі згадані, а також цілу низку інших класифікацій.

Визначимо у структурі $\sigma[\beta]$ Л-системи $V(I^{\rho}(D))$ підмножину \mathcal{A} спеціального типу, елементи A ($A \in \mathcal{A}$) якої будемо називати *автоморфізмами* Л-системи $V(I^{\rho}(D))$. Смісл цих елементів полягає в тому, що вони забезпечують внутрішні відображення $V(I^{\rho}(D))$, тобто відображення:

$$A: V(I^{\rho}(D)) \rightarrow V(I^{\rho}(D)) \quad (1.37)$$

спеціального характеру, ще точніше — відображення між окремими словниковими статтями $A: V(x) \rightarrow V(y)$ для різних x та y . У конкретних лексикографічних працях автоморфізм A може, зокрема, констатувати наявність відсилкових словникових статей типу, наприклад, таких: *x див. у*. Вказаний автоморфізм визначає таке відображення словникових статей: $V(x) \rightarrow V(y)$. Його ідентифікатором є, як правило, деяке відсилкове псевдослово (у наведеному прикладі — “*див. у*”), яке зіставляє словниковій статті $V(x)$ її відповідник $V(y)$. Але будова автоморфізма A може бути складнішою, ніж у цьому прикладі.

По-перше, довжина низки відсилань може бути більшою за одиницю, тобто мати ланцюгово чи рекурсивно розгортальний характер:

$$V(x) \rightarrow \{V(x')\} \rightarrow \dots \rightarrow \{V(x'')\} \rightarrow \dots$$

Крім того, відображення $V(x) \rightarrow V(y)$ може репрезентувати цілий пучок відсилань. Це реалізується, наприклад, тоді, коли словникова стаття $V(x)$ має таку будову: *x, x', x'', ... див. у, у', у'', ...*

У цьому випадку в одній словниковій статті $V(X)$ визначено пучок відображень:

$$V(x) \rightarrow V(y); V(x') \rightarrow V(y'); V(x'') \rightarrow V(y'') \dots$$

Наприклад, словникова стаття в 11-томному тлумачному Словнику української мови (СУМ)⁴³:

⁴³ Словник української мови в 11 томах, — К.: Наукова думка, 1970 — 1980.

УГВИНТИТИ, УГВИНТИТИСЯ, УГВИНЧЕНИЙ, УГВИНЧУВАТИСЯ *див.* вгвинтити, вгвинтитись і т.д.

репрезентує пучок відсилань:

$V(\text{УГВИНТИТИ}) \rightarrow V(\text{вгвинтити})$,

$V(\text{УГВИНТИТИСЯ}) \rightarrow V(\text{вгвинтитися})$

і т.д.

Елементи множини автоморфізмів A не обов'язково повинні задаватися в явному вигляді. Більше того, встановлення словникових автоморфізмів, як правило, не формалізується, оскільки є досить складним завданням, пов'язаним з розкриттям внутрішніх закономірностей й прихованої структури (симетрії) L -системи.

Множини відображень H та \mathcal{A} породжують *макроструктуру* L -системи. До макроструктурних елементів L -системи відносимо також $F\sigma[\beta] = \lambda[\beta]$ і $C\sigma[\beta] = \rho[\beta]$, визначені формулою (1.25). Локальні відображення (вони будуються як обмеження відповідних макроструктур на $V(x)$): $H|_{V(x)}$, а також $\lambda[\beta]|_{V(x)} \equiv \lambda(x)$; $\rho[\beta]|_{V(x)} \equiv \rho(x)$ — два останні визначено формулою (1.26) — визначають *мікроструктуру*, яка відображає у неявному вигляді семантику предметної галузі, що є об'єктом даної конкретної L -системи.

Встановлення і визначення мікроструктур лексикографічних систем дозволяє формалізувати, а у багатьох випадках й автоматизувати процес побудови структур відповідних словникових баз даних, що надає структурному підходу значні переваги при проектуванні елементів лінгвістичного забезпечення інформаційних систем.

Нижче ми вводимо деякі поняття, які дозволяють формально визначити структури, що індукуються на лексикографічних системах відображеннями спеціального вигляду. Ці відображення не тільки індукують природні структури на конкретних лексикографічних системах, але й дають певний інструментарій для породження структурної класифікації лексикографічних систем і методологію класифікації словників. З використанням цих відображень вдається сформулювати поняття близькості на множині слів (взагалі, одиниць будь-якого рівня) — ми вживаємо термін *псевдотопологія* — і навіть відстані (*псевдовідстані*) між словами.

Розглянемо деяку елементарну L -систему $V(I^W(L)) = \{V(x)\}$. Будь-яку її словникову статтю V_i завжди можна представити у вигляді:

$$V_i = x_0^i \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i, \text{ де} \quad (1.38)$$

$$\bigcup_j \xi_j^i = I^W(L), \quad (1.39)$$

де $I^W(L)$ — множина слів мови L , що містяться у всіх словникових статтях Л-системи $V(I^W(L))$; π_i — роздільники між словами (знаки пунктуації, службові позначки, спецсимволи, скорочення тощо; вони узагальнюються під назвою mark-up symbols). Отже довільна словникова стаття V_i подається у вигляді об'єднання:

$$V_i = \partial V_i \cup M_i, \quad (1.40)$$

де вжито такі позначення: $\partial V_i \equiv x_0^i$ — граничний елемент словникової статті, її заголовкове слово; M_i — “внутрішня” частина словникової статті:

$$M_i \equiv \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i. \quad (1.41)$$

Таким чином, будь-яка словникова стаття є об'єднанням внутрішньої частини та границі, а весь словник представляється у вигляді об'єднання множини внутрішніх частин з множиною границь. Позначимо:

$$\left. \begin{aligned} \partial V &= \bigcup \partial V_i \text{ — границя елементарної Л-системи } V(I^W(L)); \\ M &= \bigcup M_i \text{ — внутрішня частина елементарної Л-системи } V(I^W(L)). \end{aligned} \right\} (1.42)$$

Отже $V = \partial V \cup M$.

Визначення 2. Лексикографічна система називається замкненою, якщо для $\forall \xi \in I^W(L) \exists V(x_0^\alpha) \equiv V_\alpha \in V$, що $x_0^\alpha \equiv x\xi$, де $x\xi$ — вихідна форма слова ξ .

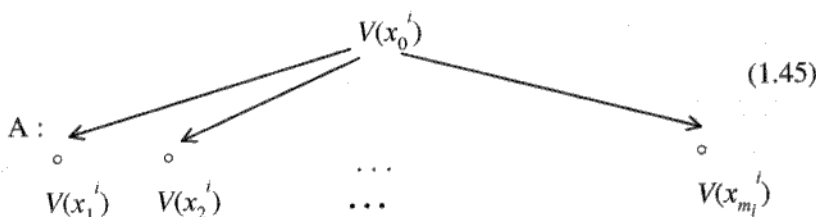
Визначимо автоморфізм A деякого спеціального типу в такий спосіб. Назвемо множину слів:

$$(x_1^i, x_2^i, \dots, x_{m_i}^i); m_i \leq n_i \quad (1.43)$$

кортежем словникової статті V_i , якщо вона являє собою множину канонічних форм до слів $\xi_1^i, \xi_2^i, \dots, \xi_{m_i}^i$, відповідно. До кортежу необов'язково включаються всі слова з M_i . При розгляді конкретних прикладів, зокрема, при вивченні структури лексикографічних систем тлумачного типу, лексикографічна доцільність спонукає до обмеження кортежу лексемами, що входять тільки до формул тлумачення (словникових дефініцій), причому з випущенням з них так званих стоп-слів (таких, що не дають “суттєвого” внеску в семантику). Визначимо автоморфізм $A \in \mathcal{A}$ за допомогою формули:

$$A : V(x_0^i) \equiv V_i \rightarrow \{V(x_k^i), k = 1, 2, \dots, m_i\} \quad (1.44)$$

Графічна репрезентація формули (1.44) має такий вигляд:



Формули (1.44)—(1.45) означають, що словниковій статті $V(x_0^i)$ із заголовковим словом (“границею”) x_0 ставляться у відповідність словникові статті $V(x_k^i)$, $k = 1, 2, \dots, m_i$ із заголовковими словами x_k^i , $k = 1, 2, \dots, m_i$, відповідно, якщо, зрозуміло, вони взагалі наявні у $V(I^W(L))$. Рекурсивно довизначимо дію оператора A на $V(x_1^i)$, $V(x_2^i)$, ..., $V(x_{m_i}^i)$, і далі — на результатах його застосування до $V(x_1^i)$, $V(x_2^i)$, ..., $V(x_{m_i}^i)$ і т. д. аж допоки об’єкти $V(x_j^i)$, $i, j = 1, 2, \dots$, не почнуть повторюватися. Через $V^A[x_0^i]$ ($V^A[x_0^i] \subseteq V$) позначимо множину словникових статей $\{V(x_0^i), V(x_1^i), V(x_2^i), \dots, V(x_{m_i}^i), \dots\}$, одержану в результаті визначеної вище дії оператора A .

Визначення 3. Множина $V^A[x_0^i]$ називається $A[x_0^i]$ -підсловником словника $V(I^W(L))$, якщо $A V^A[x_0^i] = V^A[x_0^i]$.

Таким чином A -підсловник $V^A[x_0^i]$ є інваріантною множиною у $V(I^W(L))$ відносно дії A .

Визначення 4. Елементи $V(x_0^i)$, $V(x_1^i)$, $V(x_2^i)$, ..., $V(x_{m_i}^i)$, ..., A -підсловника $V^A[x_0^i]$ називатимемо A -еквівалентними елементами.

Останнє визначення стає зрозумілішим, якщо врахувати, що множина відображень A , яка породжує A -підсловник $V^A[x_0^i]$, індукує відношення еквівалентності на множині його словникових статей; його позначатимемо символом $EV^A[x_0^i]$. Позначимо через:

$$W = V \setminus EV^A[x_0^i] \quad (1.45)$$

фактормножину в V по відношенню $EV^A[x_0^i]$. Звідси: $V = W \cup V'$. Згідно з визначенням V' :

$$A V' = V' A^M W = V' \text{ для деякого } M \geq 0.$$

Тоді говоримо, що V являє собою *напівпряму суму* словників W і V' :

$$V = W \triangleright V'. \quad (1.46)$$

Словник V з такою структурою будемо називати A -нерозкладним.

Визначення 6. Словник $V(I^W(L))$ називається A -незвідним (цілком незвідним), якщо в ньому немає власних A -підсловників.

З останнього визначення випливає, що якщо V є A -незвідним словником, то для довільних $x, y \in I^W(L) \exists N \geq 0$, що $V(y) \subseteq A^N V(x)$.

Визначення 7. Словник V називається цілком A -звідним, якщо його можна подати у вигляді:

$$V = \bigcup_i V^i, \text{ причому } V^i \cap V^j = \emptyset \text{ при } i \neq j,$$

де V^i — A -незвідний словник. У цьому випадку будемо говорити, що V розкладається в пряму суму A -словників V^i :

$$V = \sum_i \oplus V^i. \quad (1.47)$$

Розглянемо деякий A -незвідний словник V :

$$V = \bigcup_{x_0 \in S_0} V(x_0^i),$$

де $V(x_0^i)$ — словникова стаття з заголовковим словом x_0^i . Авто-морфізм A індукує відображення $S_0 \rightarrow S_0$, тобто він визначає деяке відображення множини заголовкових слів *в себе*. Вказане відображення визначається у такий спосіб: будемо вважати, що якщо:

$$A: V(x_0^i) \equiv V_i \rightarrow \{V(x_k^i), k = 1, 2, \dots, m_i\}, \text{ то}$$

$$A: x_0^i \rightarrow \{x_k^i, k = 1, 2, \dots, m_i\}.$$

З викладеного випливає

Теорема 1. Для A -незвідного словника V не існує A -інваріантних підмножин в S_0 . Отже для довільних $x, y \in S_0 \exists N(x, y) \geq 0$ таке, що $A^N x = y$.

Іншими словами, шлях $A^N x$ при достатньо великих N проходить через всі точки множини S_0 . Як наслідок справедлива

Теорема 2. Повний A -шлях на графі $G^A(S_0)$ завжди замкнений.

Позначимо: $\inf N(x, y) = \rho(x, y)$.

Визначення 8. Число $\rho(x, y)$ називається A -псевдовідстанню від слова x до слова y .

Число $\rho(x, y)$ показує, за яку мінімальну кількість кроків можна дійти від слова x до слова y , застосовуючи алгоритм A .

Нехай множина слів:

$$(x_1^i, x_2^i, \dots, x_{m_i}^i; m_i \leq n_i) = \tau^i \equiv \tau(x_0^i)$$

є кортежем словникової статті V_{i, p_i} , тобто являє собою множину вихідних форм до слів $\xi_1^i, \xi_2^i, \dots, \xi_{m_i}^i$, відповідно. Очевидно, що x_0^i також $\in \tau(x_0^i)$.

Визначення 9. Назвемо $\tau(x_0^i)$ замкненим околom точки x_0^i .

Визначення 10. Множина

$$\{\emptyset, \tau(x_0^i), i = 1, 2, \dots, \text{Card } S_0\} \quad (1.48)$$

називається (V, A) -псевдотопологією лексикографічної системи.

Поняття псевдотопології спроможне відіграти важливу роль у теорії лексикографічних систем, оскільки за допомогою цього поняття виникає можливість формалізації поняття близькості елементарних інформаційних одиниць (слів). А саме, перетин околів $\tau(x_0^i) \cap \tau(x_0^j)$ визначає ступінь близькості лексем x_0^i та x_0^j .

1.4.4. Лексикографічні середовища

У реальній дійсності об'єкти мови функціонують у їх цілісності, не поділеній на окремі складники концептуального представлення. У лексикографічній системі це виявляється в тому, що при моделюванні мовних об'єктів засобами теорії Л-систем постає завдання інтеграції різних типів лексикографічних ефектів, а також поєднання й узгодження різнорідних (гетерогенних) лексикографічних структур. У свою чергу, це потребує узгодження між усіма елементами архітектури Л-систем, які підлягають процесу інтеграції.

Численні експерименти зі створення конкретних комп'ютерних реалізацій інтегрованих лінгвістичних об'єктів дозволили зробити висновок про необхідність побудови спеціального мовно-інформаційного середовища, яке б із самого початку було адаптованим до процесів інтегрування різних лексикографічних систем і містило необхідні засоби та конструктиви для здійснення зазначених процесів, а також фіксації їх результатів у вигляді інтегрованих лексикографічних систем, що мають відмінні лексикографічні

структури. Як наслідок, було запропоновано поняття лексикографічного середовища⁴⁴.

Визначення. Вважається, що задано лексикографічне середовище (L -середовище) ML , якщо:

1. Задано клас $Ob ML$ елементів, кожен з яких є діаграмою вигляду (1.36) і представляє певну L -систему (не обов'язково елементарну). Елементи з $Ob ML$ називаються об'єктами L -середовища ML — позначатимемо їх великими латинськими літерами: A, B, C, \dots .

2. Для кожної пари об'єктів A, B з ML задано множину $Hom_{ML}(A, B)$, яка називається множиною морфізмів A в B ; замість $f \in Hom_{ML}(A, B)$ також пишуть:

$f: A \rightarrow B$ або $A \xrightarrow{f} B$. При цьому $f: CM_A \rightarrow CM_B; f: INM_A \rightarrow INM_B;$

$f: EXM_A \rightarrow EXM_B; f(\varphi_A) = \varphi_B; f(\psi_A) = \psi_B; f(\zeta_A) = \zeta_B$ та $f(\xi_A) \circ f(\psi_A) = f(\varphi_A)$.

3. Для кожної трійки об'єктів (A, B, C) з ML задано відображення

$\mu: Hom_{ML}(A, B) \times Hom_{ML}(B, C) \rightarrow Hom_{ML}(A, C)$

(образ $\mu(f, g)$ пари (f, g) , де $f \in Hom_{ML}(A, B)$, $g \in Hom_{ML}(B, C)$), буде позначатися $f \circ g$ або $f g$ і називатися композицією морфізмів f і g .

4. Множини $Hom_{ML}(A, B)$ і композиція морфізмів задовольняють таким аксіомам:

(а) Асоціативність: для кожної трійки морфізмів f, g, h :

$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D \quad f(g h) = (f g) h.$

(б) Існування одиниці: для кожного $A \in Ob ML$ існує морфізм $1_A: A \rightarrow A$, ($1_A \in Hom_{ML}(A, A)$), такий що $1_A f = f$ і $g 1_A = g$ для довільних морфізмів $f \in Hom_{ML}(B, A)$ і $g \in Hom_{ML}(A, B)$.

(с) Якщо пари (A, B) і (A', B') різні, перетин $Hom_{ML}(A, B)$ і $Hom_{ML}(A', B')$ порожній.

Нехай дано два лексикографічні середовища ML_1 і ML_2 . Коваріантний (відповідно контраваріантний) функтор F з ML_1 в ML_2 складається з:

(а) відображення $A \rightarrow F(A)$, яке зіставляє кожен об'єкт $A \in Ob ML_1$ з об'єктом $F(A) \in Ob ML_2$;

(б) відображень $F(A, B): Hom_{ML_1}(A, B) \rightarrow Hom_{ML_2}(F(A), F(B))$ — для коваріантного й $F(A, B): Hom_{ML_1}(A, B) \rightarrow Hom_{ML_2}(F(B), F(A))$ —

⁴⁴ Рабуць О. Г. Інтегровані лексикографічні системи: Автореф. дис. ... канд. техн. наук. — К., 2002. — 18 с.

для контраваріантного функторів, визначених для всіх пар (A, B) об'єктів з \mathbf{ML}_1 і таких, що (якщо замість $F(A, B)(u)$ писати $F(u)$) $F(1_A) = 1_{F(A)}$ і $F(vu) = F(v)F(u)$ (відповідно $F(vu) = F(u)F(v)$).

1.4.5. Інтегровані Л-системи та методика їх побудови

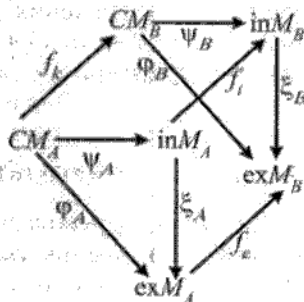
Уведена в описаний вище спосіб структура Л-середовища є зручним формальним об'єктом для формування складних лексикографічних конструкцій, які поєднують в одне ціле багато окремих різномірних (гетерогенних) Л-систем. Гетерогенність Л-систем, які підлягають інтеграції, є багатоаспектним поняттям. Ми розглядаємо Л-системи, неоднорідні за всіма рівнями архітектури — концептуальним, внутрішнім і зовнішнім. Такий підхід передбачає розроблення методів інтеграції концептуальних моделей, способів представлення даних та операційно-програмних платформ, а також узгодження зовнішніх представлень відповідних концептуальних схем та їх внутрішніх репрезентацій.

Під інтеграцією Л-систем як інформаційних систем розуміємо досягнення можливості одночасного та спільного використання прикладною програмою кількох інформаційних систем як єдиного цілого. Аналогічні визначення пропонує Державний стандарт України ДСТУ 2941—94: інтегрована система — сукупність двох або кількох взаємопов'язаних систем, у якій функціонування однієї з них залежить від результатів функціонування іншої (інших) так, що цю сукупність можна розглядати як єдину систему; інтегрування систем — об'єднання кількох систем різного призначення в єдину багатофункціональну систему.

Отже, для прикладної програми інтегрована сукупність різних лексикографічних баз даних (ЛБД) повинна мати вигляд як єдина ЛБД. Ці уявлення певною мірою можуть бути перенесені і на традиційні словники, які також можуть являти собою інтегровані Л-системи, зокрема, тлумачний словник є одним із найяскравіших прикладів такої Л-системи.

У цьому підрозділі максимальний акцент зроблено на інтеграції Л-систем на концептуальному рівні. Базовими об'єктами, що підлягають інтеграції, є діаграми вигляду (1.36), які ми розглядаємо як об'єкти певного Л-середовища.

Нехай є два об'єкти A і B , а також $f: A \rightarrow B, f \in \text{Hom}_{\mathbf{ML}}(A, B)$. Побудуємо f як спеціальний морфізм у вигляді трикомпонентного вектора: $f = (f_a, f_b, f_c)$ такого, що діаграма:



(1.49)

є комутативною — це еквівалентно справедливості таких рівностей:

$$\psi_B \circ f_k = f_l \circ \psi_A; \varphi_B \circ f_k = f_e \circ \varphi_A; \xi_B \circ f_i = f_e \circ \xi_A \quad (1.50)$$

Діаграма (1.49) та рівності (1.50) надають необхідні формальні засоби, мовою яких можуть бути сформульовані процеси побудови інтеграційної архітектури, що впливає з наступного викладу.

Морфізм $f = (f_k, f_l, f_e)$, $f: A \rightarrow B$ називається регулярним, якщо він задовольняє умовам:

— повної визначеності, згідно з якою будь-якому стану A відповідає один і тільки один стан B ;

— інтерпретованості, згідно з якою будь-якому елементу, який належить до множини $\{\sigma_A[\beta_A], RR\downarrow[V(A)]\}$, однозначно відповідає елемент з множини $\{\sigma_B[\beta_B], RR\downarrow[V(B)]\}$;

— відтворюваності, згідно з якою зміні будь-якого стану A , що здійснюється певним оператором з $\{\sigma_A[\beta_A], RR\downarrow[V(A)]\}$, відповідає адекватна зміна відповідного стану, що виконується певним оператором з $\{\sigma_B[\beta_B], RR\downarrow[V(B)]\}$.

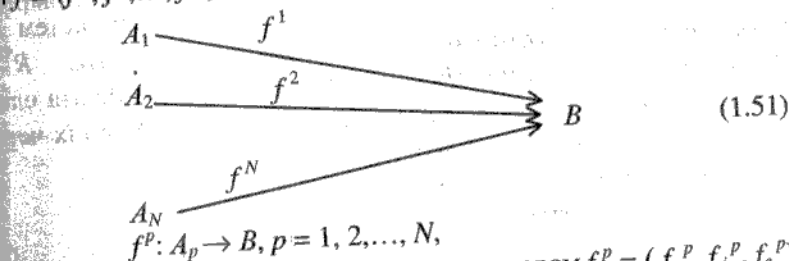
Поряд зі здатністю до інтеграції концептуально гетерогенних лексикографічних систем, що репрезентують різні мовні явища, не менш важливим аспектом інтеграційної архітектури є можливість досягнення високого ступеня незалежності прикладних програм від СКБД і забезпечення їхньої мобільності щодо СКБД різних типів. Проблема мобільності програм узагалі формулюється як забезпечення можливості виконання певної програми на різних комп'ютерних платформах без її зміни. Аналогічно визначається питання щодо мобільності програм стосовно СКБД.

Теоретично для цього існують такі можливості: 1) створення універсальної мови програмування й вимога її всезагального застосування; 2) забезпечення кожного комп'ютера компіляторами для

всіх мов програмування при належній їх стандартизації; 3) введення платформи-незалежної проміжної мови з її інструментальною реалізацією на рівні віртуальної машини з вбудованими інтерфейсами до будь-яких платформ; 4) застосування емуляційних методів; 5) використання комп'ютерних мереж за припущення, що хоча б один із мережевих комп'ютерів забезпечено необхідним компілятором.

Аналіз викладених пропозицій приводить до висновку про раціональність саме третьої з них. У різних варіаціях її намагаються реалізувати, наприклад, у системах типу JAVA, стандартизації машинно-незалежних мов SQL, SQL2 тощо. Зауважимо: зазначені пропозиції зовні аналогічні конструюванню спільної концептуальної моделі, що вкотре підтверджує положення про центральність саме концептуальної моделі в архітектурі інформаційної системи. Тому далі розглядатимемо саме розробку процесів інтеграції Л-систем на концептуальному рівні (або інтеграції концептуальних представлень Л-систем).

Зміст процедури інтеграції полягає ось у чому. Припустимо, що замість одного об'єкта A , наявного в діаграмі (1.49), маємо набір об'єктів A_1, A_2, \dots, A_N . Збудуємо віялоподібне відображення з морфізмами $f = (f^1, f^2, \dots, f^N)$:



де кожний f^p задається трикомпонентним вектором $f^p = (f_c^p, f_i^p, f_e^p)$ з властивостями (1.50).

У процесі застосування даних морфізмів використовується інтеграційна процедура, яка має такі етапи. Спочатку будується допоміжна Л-система B з таких елементів: знакові системи $A(A_p)$; структури $\beta_p, \sigma_p[\beta_p]$ та процеси $RR_p \downarrow [V(I^Q(S))]$ всіх A_p . На цьому етапі відображення $f^p: A_p \rightarrow B, p = 1, 2, \dots, N$ інтерпретується як відображення вкладення. Так отримуємо певну неелементарну Л-систему B з незалежними складниками $A_p, p = 1, 2, \dots, N$.

Знакова система для Л-системи B будується як об'єднання: $A(B) = \cup A(A_p)$.

Далі вводимо спеціальну семантичну процедуру SEM, яка забезпечує ототожнення елементів структури, що збігаються бодай для двох систем з множини A_p , $p = 1, 2, \dots, N$. Вважатимемо, що ця процедура здійснює ототожнення не лише імен семантично тотожних атрибутів, наявних у різних Л-системах A_p , $p = 1, 2, \dots, N$, а й відповідних областей їхніх значень (доменів). При цьому можливі різні випадки перетинів структур, що належать до різних A_p , $p = 1, 2, \dots, N$. Розглянемо їх докладніше.

Позначимо через $\epsilon_{p_1 p_2 \dots p_k}[\beta]$ множини елементів структури (разом з їхніми доменами), які належать водночас до кожного з $A_{p_1}, A_{p_2}, \dots, A_{p_k}$, $1 \leq p_1 < p_2 < \dots < p_k \leq N$. Застосуємо до неї семантичну операцію:

$$\text{SEM}(\epsilon_{p_1 p_2 \dots p_k}[\beta]) = \epsilon_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta] \quad (1.52)$$

і в такий спосіб одержимо семантично тотожні спільні елементи структури Л-систем з номерами p_1, p_2, \dots, p_k . Наявність елементів (1.19) дає можливість побудови Л-системи зі структурою:

$$\sigma_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta] = (\epsilon_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta], [b_{p_1}^{\text{SEM}}], [b_{p_2}^{\text{SEM}}], \dots, [b_{p_k}^{\text{SEM}}]; R^{\text{SEM}}), \quad (1.53)$$

де $[b_i^{\text{SEM}}]$, $i = p_1, p_2, \dots, p_k$ — елементи структури Л-систем A_i , в яких виконано ототожнення спільних елементів; R^{SEM} — об'єднання операцій, що діють у кожній Л-системі. Решта елементів структури Л-систем A_i , $i = p_1, p_2, \dots, p_k$, — позначимо їх через S_i , $i = p_1, p_2, \dots, p_k$, залишається без зміни.

У такий спосіб Л-система \underline{B} зі структурою:

$$\{\sigma_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta]; S_i, i = p_1, p_2, \dots, p_k\} \quad (1.54)$$

інтегрує Л-системи A_i , $i = p_1, p_2, \dots, p_k$.

Позначимо через $\epsilon[\beta]$ множини елементів структури (разом з їх доменами, які належать до всіх A_p , $p = 1, 2, \dots, N$):

$$\epsilon[\beta] = \bigcap_p^N [\beta] A_p. \quad (1.55)$$

У результаті одержуємо Л-систему \underline{B} зі знаковою системою $A(\underline{B}) = \cup A(A_p)$ та структурою $\beta(\underline{B})$, $\sigma[\beta](\underline{B})$, $RR \downarrow [V(\underline{B})]$, одержаною описаним вище способом. Особливо цікавим є випадок, коли певний елемент або певні елементи структури наявні в усіх Л-сис-

темах, що підлягають інтеграції, тобто коли $\varepsilon[\beta] \neq \emptyset$. Тоді доцільно провести суцільне індексування всіх A_p за значеннями відповідних доменів, які належать до елементів $\varepsilon[\beta]$. Ці елементи структури набувають статусу вхідних до кожної з A_p . Як показує досвід, елементами структури відзначеного типу для природномовних систем виступають лексичні масиви (множини словоформ) усіх мов, що беруть участь в A_p . Отже, постає завдання побудови процедури природномовної індексації.

Побудована описаним способом L -система B називається інтеграцією L -систем A_p , $p = 1, 2, \dots, N$.

Структура лексикографічного середовища містить набір елементів, необхідний для представлення ефектів граматичної та лексичної семантики. Зокрема, це стосується і подання окремих відношень і мовних явищ, які абстрагуються з мовного континууму — словозміна, словотвір, орфоєпія, фразеологія, синонімія, антонімія тощо. Більше того, конструкція лексикографічного середовища допускає моделювання кожного із зазначених явищ та відношень окремо з наступною інтеграцією їх в єдиний лексикографічний комплекс. Приклади застосування викладеної теорії при побудові конкретних лексикографічних середовищ для експлікації властивостей української лексики викладено у наступних розділах.

ЕЛЕКТРОННІ ГРАМАТИЧНІ СЛОВНИКИ

2.1. Моделювання словозмінної системи

2.1.1. Морфологічний аналіз

Перейдемо у цьому та наступних розділах до впровадження розвинутих вище формалізованих лексикографічних схем, тобто їх застосування до конкретних явищ та структур природної мови.

Першим об'єктом у цьому колі є явище словозміни, яке має велике і теоретичне — з огляду на повноту опису мовної системи та проблематику граматичної ідентифікації одиниць мови — і, передусім, практичне значення.

Для мов флективного типу, які мають багату словозмінну парадигму, серед задач граматичної ідентифікації складним і працемістким завданням є формалізація процесів лексичної парадигматизації та лематизації, а саме отримання повного списку словоформ певної лексики в усіх граматичних значеннях і виведення вихідної словоформи з будь-якої текстової словоформи. Таке завдання вирішується створенням граматичного словника, який становить основу граматичної (морфологічної) підсистеми практично кожної системи опрацювання природної мови (ОПМ).

Системно повний граматичний словник повинен для кожного слова вказувати всі ті його характеристики, які є істотними для побудови граматично правильних фраз, до складу яких входить відповідне слово. Першим етапом його створення є побудова словозмінного словника. Прикладом словозмінного словника для слов'янських мов є граматичний словник російської мови А. Залізняка¹. Зазначений словник є традиційним лексикографічним твором, виданим у паперовій формі. На його основі для російської мови було розроблено чимало граматичних комп'ютерних систем та автоматичних морфологічних аналізаторів. Для української мови такого словника укладено не було, тому граматична підсистема в УМІФі із самого початку створювалася як комп'ютерна лексикографічна система.

¹ Залізник А.А. Грамматический словарь русского языка: Словоизменение. 3-е изд. стереотип. — М.: Рус. яз., 1987. — 879 с.

Згідно з теорією та технологією лексикографічних систем граматичний словник української мови реалізовано у вигляді інструментального комплексу, який поєднує довідкову функцію, апарат для ведення, актуалізації і розширення граматичної бази даних на основі засад інформаційно-лексикографічного моделювання. Концептуальною основою такого комплексу є словозмінна класифікація української лексики².

Отже, для побудови граматичного словника визначальним фактором є наявність формальної моделі словозміни, що означає установаження та формалізацію лінгвістичних критеріїв, згідно з якими вся множина слів мови розбивається на певні підмножини, взаємний перетин яких є порожнім і в середині якого словозміна відбувається за єдиним правилом (алгоритмом). Завдання полягає у визначенні формальних ознак, за допомогою яких весь масив української лексики розподіляється на підмножини слів саме з такими властивостями — вони одержали назву словозмінних парадигматичних класів (СПК).

Морфологічні модулі, розроблені в УМІФі, реалізують алгоритми аналізу, лематизації та синтезу текстових одиниць, які передбачають використання електронного граматичного словника (ЕГС) тієї мови, якою написаний текст, що аналізується (обробляється).

Зазначимо, що існують й інші підходи до побудови морфологічних модулів: наприклад, імовірнісний підхід (стеммінг)³, а також підхід, який ґрунтується на побудові емпіричних формул⁴. Згадані підходи хоча і більш прості й певною мірою більш гнучкі, але, як відзначають самі автори, мають великий недолік — більш низьку точність, ніж у традиційних методів, які ґрунтуються на великих морфологічних словниках із застосуванням граматичних/морфологічних правил. Крім того, на нашу думку, такі підходи придатні не для будь-якої мови, а тільки для мови із простою системою словозміни (наприклад, англійської). Для мов флективних найбільш доцільно все-таки використовувати алгоритми (аналізу, синтезу, лематизації), які будуються на використанні електронних граматич-

² Шевченко І.В. Алгоритмічна словозмінна класифікація української лексики. // Мовознавство. — 1996. — № 4—5. — С. 40-44.

³ А. Коваленко. Вероятностный морфологический анализатор (стеммер): <http://linguist.nm.ru/ling/>.

Також: Porter M. An Algorithm for Suffix Stripping. //Program. 1980. #14. P. 130-137.

⁴ П. Макарашов, М. Александров, А. Гельбух. Формулы проверки подобия слов с обучением на примерах: построение и применение. //Труды Международной конференции "Корпусная лингвистика — 2004". Санкт-Петербург. С. 239-255.

них словників, і тут покладаємося саме на цей принцип. Тому спочатку подамо докладний виклад формальних та технологічних засад створення ЕГС, а потім наведемо алгоритми морфологічного аналізу та синтезу, які базуються на застосуванні електронних граматичних словників.

2.1.2. Формальні засади побудови ЕГС

(Морфологічна модель словозміни флективної мови)

Уведемо позначення. Нехай L — є певною (флективною) мовою⁵, W — множина слів мови L , t_i , ($i = 1, 2, \dots, N$) — морфологічні типи, $W(t_i)$ — множина слів мови L , яка належить типу t_i , $\Omega(t_i)$ — множина граматичних значень, що відповідають типу t_i .

Парадигматичний тип, що характеризується граматичними формами, які визначаються граматичними значеннями словозмінних категорій “число” та “відмінок” будемо називати субстантивним парадигматичним типом⁶:

$$W(t_i) \equiv W^S = \{w^S_1, w^S_2, \dots, w^S_{12}, w^S_{14}\}. \quad (2.1)$$

При цьому граматичні форми $w^S_i = w^S_i(n, k)$,⁷ визначаються множинами/парами (число, відмінок):

$$\begin{aligned} w^S_i &= \{n_1, k_i\} = \{(n_1, k_1), (n_1, k_2), (n_1, k_3), (n_1, k_4), (n_1, k_5), (n_1, k_6), (n_1, k_7)\}, \\ w^S_{i+7} &= \{n_2, k_i\} = \\ &= \{(n_2, k_1), (n_2, k_2), (n_2, k_3), (n_2, k_4), (n_2, k_5), (n_2, k_6), (n_2, k_7)\}, \end{aligned} \quad (2.2)$$

$i = 1, 2, \dots, 6, 7$,

де n_1 — одна, n_2 — множина; k_i — значення відмінків: k_1 — називний, k_2 — родовий, k_3 — давальний, k_4 — знахідний, k_5 — орудний, k_6 — місцевий, k_7 — кличний.

До субстантивного парадигматичного типу належать всі іменники і займенники-іменники (займенники, які є заміниками іменників — особові займенники: *я, ти, він, ...*; та такі як *хто, що, де-хто, ...*).

⁵ Виклад наводиться на прикладі української, а також для російської мови, виклад для якої є аналогічним.

⁶ Для російської мови субстантивний парадигматичний тип характеризується 12 граматичними формами: 6 відмінків у однині + 6 відмінків у множині. Для української мови субстантивний парадигматичний тип характеризується 14 граматичними формами внаслідок того, що відмінків в українській граматиці 7, а не 6 як в російській.

⁷ Граматична форма залежить від значень категорій *число* та *відмінок*.

Відзначимо, що в кожній з граматичних форм конкретна лексема може мати одну або декілька словоформ, чи не мати жодної словоформи. У випадку відсутності реалізації лексики в певній (конкретній) граматичній формі будемо говорити про дефектність словозмінної парадигми. Прикладом подібних випадків є відсутність форм множини у іменників *singularia tantum*, або відсутність форм однини у іменників *pluralia tantum*. Для урахування факту відсутності словоформ у деяких граматичних формах у формулу опису парадигматичного типу (2.1) введемо параметр *def*, який будемо називати параметром дефектності:

$$W^S = \{w^S_1, w^S_2, \dots, w^S_{14}, def\}, \quad (2.3)$$

що вказує на номери (числа) граматичних форм, для яких відповідні словоформи відсутні в повній парадигмі; якщо дефектності немає, то за визначенням покладаємо *def* = 0.

Парадигматичний тип, що характеризується граматичними формами, які визначаються граматичними значеннями словозмінних категорій "рід", "число" та "відмінок" будемо називати ад'єктивним парадигматичним типом⁸:

$$W(t_2) \equiv W^A = \{w^A_1, w^A_2, \dots, w^A_{28}\}, \quad (2.4)$$

де граматичні форми w_i представлені трійками (рід, число, відмінок); для форм однини:

$$w^A_i = \{g_1, n_1, k_1\}, w^A_{i+6} = \{g_2, n_1, k_1\}, w^A_{i+12} = \{g_3, n_1, k_1\}, i = 1, 2, \dots, 6, \quad (2.5)$$

а для форм множини значення роду нерелевантне:

$$w^A_{i+18} = \{n_2, k_1\}, i = 1, 2, \dots, 6. \quad (2.6)$$

У формулах (2.4)—(2.6) g_1 — чоловічий рід, g_2 — жіночий рід, g_3 — середній рід, n_1 — однина, n_2 — множина; k_1 — значення відмінків: k_1 — називний, k_2 — родовий, k_3 — давальний, k_4 — знахідний, k_5 — орудний, k_6 — місцевий;

короткі граматичні форми w^A_{25} , w^A_{26} , w^A_{27} , w^A_{28} існують тільки для називного відмінку і визначаються категоріями роду та числа:

$$\begin{aligned} w^A_{25} &= \{g_1, n_1, k_1\}, w^A_{26} = \{g_2, n_1, k_1\}, w^A_{27} = \\ &= \{g_3, n_1, k_1\}, w^A_{28} = \{n_2, k_1\}. \end{aligned} \quad (2.7)$$

⁸ Ад'єктивний парадигматичний тип української та російської мов характеризується 28 граматичними значеннями: 3 значення роду x 6 відмінків однини + 6 відмінків множини. Для російської мови слід додати ще 4 коротких форми (м.р., ж.р., с.р. і мн.ч.). В українській мові для ад'єктивного парадигматичного типу коротких форм немає.

До ад'єктивного парадигматичного типу належать прикметники, займенники-прикметники, порядкові числівники, дісприкметники, кількісний числівник "один".

В ад'єктивному парадигматичному типі, так само як і в субстантивному, парадигма може бути дефектною. Формула опису парадигматичного типу з урахуванням параметра дефектності має вигляд:

$$W^A = \{w^A_1, w^A_2, \dots, w^A_{24}, def\}, \text{ для української} \\ \text{та } W^A = \{w^A_1, w^A_2, \dots, w^A_{28}, def\}, \text{ для російської} \quad (2.8)$$

де w_i визначаються за формулами (2.5) — (2.8), а параметр def має той же зміст, що і в (2.3).

Словозміна дієслів характеризується граматичними формами, що визначаються граматичними значеннями категорій "стан", "час", "число", "особа", "спосіб", "рід" ("залог", "время", "число", "лицо", "наклонение", "род") (категорія роду релевантна тільки для минулого часу). Отже дієслівний парадигматичний тип описується формулою:

$$W(t_3) \equiv W^V = \{w^V_0, w^V_1, \dots, w^V_{43}, def\}, \quad (2.9)$$

де w_0 — інфінітив дієслова (збігається з реєстровим словом); $\{w^V_1, w^V_2, \dots, w^V_6\}$ — представляють граматичні форми дійсного стану теперішнього часу (дійствительного залога настоящего времени); w^V_i ($i=1,2,\dots,6$) визначаються п'ятірками категорій (стан, спосіб, час, число, особа)⁹ при фіксованих значеннях стану ($z=z_1$ — активний стан), способу ($h=h_1$ — дійсний спосіб) и часу ($t=t_1$ — теперішній час):

$$w^V_i = \{z_1, h_1, t_1, n_1, l_i\}, w^V_{i+3} = \{z_1, h_1, t_1, n_2, l_i\}, i=1,2,3. \quad (2.10)$$

$\{w^V_7, w^V_8, \dots, w^V_{10}\}$ — граматичні форми активного стану минулого часу; w^V_{i+6} ($i=7,8,9,10$) визначаються п'ятірками категорій (стан, спосіб, час, число, рід)¹⁰ при фіксованих значеннях стану ($z=z_1$), способу ($h=h_1$) и часу ($t=t_2$ — минулий час):

$$w^V_{i+6} = \{z_1, h_1, t_2, n_1, g_i\}, i=1,2,3; \quad (2.11)$$

$$w^V_{10} = \{z_1, h_1, t_2, n_2\}. \quad (2.12)$$

$\{w^V_{11}, w^V_{12}, \dots, w^V_{16}\}$ — граматичні форми активного стану (дійствительного залога) майбутнього часу; w^V_{i+11} ($i=11,12,\dots,16$) визнача-

⁹ Категорія роду не є релевантною для форм теперішнього і майбутнього часу.

¹⁰ Для форм минулого часу категорія особи не є релевантною.

ються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z=z_1$), способу ($h=h_1$) і часу ($t=t_3$ — майбутній час):

$$w_{i+10}^V = \{z_1, h_1, t_3, n_1, l_i\}, w_{i+12}^V = \{z_1, h_1, t_3, n_2, l_i\}, i=1,2,3. \quad (2.13)$$

$\{w_{17}^V, w_{18}^V\}$ — граматичні форми наказового способу, які визначаються четвірками категорій (*стан, спосіб, число, особа*) при фіксованих значеннях стану ($z=z_1$), способу ($h=h_3$ — наказовий спосіб) та особи ($l=l_2$ — друга особа):

$$w_{i+16}^V = \{z_1, h_3, n_i, l_2\}, i=1,2; \quad (2.14)$$

$\{w_{19}^V, w_{20}^V\}$ — дієприслівникові (деєпричастные) граматичні форми, які визначаються значеннями пар категорій (*стан, час*) при фіксованих значеннях стану ($z=z_1$) (дієприслівники активного стану теперішнього та минулого часу):

$$w_{i+18}^V = \{z_1, t_i\}, i=1,2; \quad (2.15)$$

$\{w_{21}^V, w_{22}^V, w_{23}^V, w_{24}^V\}$ — дієприкметникові (причастные) граматичні форми дієслова, які визначаються категоріями (*стан, час*) (дієприкметники активного та пасивного стану теперішнього та минулого часу):

$$w_{i+20}^V = \{z_1, t_i\}, w_{i+20}^V = \{z_2, t_i\}, i=1,2; \quad (2.16)$$

w_{25}^V — інфінітив пасивної форми дієслова;

$\{w_{26}^V, w_{27}^V, \dots, w_{31}^V\}$ — граматичні форми пасивного стану теперішнього часу; значення w_i^V ($i=26,27,\dots,31$) визначаються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z=z_2$ — пасивний стан), способу ($h=h_1$ — дійсний спосіб) та часу ($t=t_1$ — теперішній час):

$$w_{i+25}^V = \{z_2, h_1, t_1, n_1, l_i\}, w_{i+28}^V = \{z_2, h_1, t_1, n_2, l_i\}, i=1,2,3; \quad (2.17)$$

$\{w_{32}^V, w_{33}^V, \dots, w_{35}^V\}$ — граматичні форми, що визначаються п'ятірками категорій (*стан, спосіб, час, число, рід*) при фіксованих значеннях стану ($z=z_2$), способу ($h=h_1$) і часу ($t=t_2$ — минулий час):

$$w_{i+31}^V = \{z_2, h_1, t_2, n_1, g_i\}, i=1,2,3; \quad (2.18)$$

$$w_{35}^V = \{z_2, h_1, t_2, n_2\}, \quad (2.19)$$

$\{w_{36}^V, w_{37}^V, \dots, w_{41}^V\}$ — граматичні форми пасивного стану майбутнього часу; значення w_i^V ($i=36,37,\dots,41$) визначаються п'ятірками категорій (*стан, спосіб, час, число, особа*) при фіксованих значеннях стану ($z=z_2$ — пасивний стан), способу ($h=h_1$ — дійсний спосіб) і часу ($t=t_3$ — майбутній час):

$$w_{i35}^V = \{z_2, h_1, t_3, n_1, l_i\}, w_{i38}^V = \{z_2, h_1, t_3, n_2, l_i\}, i=1,2,3. \quad (2.20)$$

$\{w_{42}^V, w_{43}^V\}$ — граматичні форми пасивного стану наказового способу, що визначаються четвірками категорій (*стан, спосіб, число, особа*) при фіксованих значеннях стану ($z=z_1$), способу ($h=h_3$ — наказовий спосіб) та особи ($l=l_2$ — 2-а особа):

$$w_{i41}^V = \{z_2, h_3, n_i, l_2\}, i=1,2; \quad (2.21)$$

def — параметр дефектності.

Формули (2.10)—(2.21) описують всі основні можливі граматичні форми синтетичних дієслівних форм, які можуть бути властиві парадигмі дієслова (і які звичайно залучаються розробниками до словозмінної парадигми дієслова). Не включено аналітичні форми, зокрема форми умовного способу, а також зворотні форми дієслова. Зворотнє дієслово розглядається нами як самостійне і змінне згідно парадигматичного типу, що описується формулами (2.9)—(2.21). Для парадигматичного типу дієслів можливі випадки дефектності декількох видів. Всі вони описуються параметром *def*, який вказує номери граматичних форм, для котрих відсутні варіанти слів-форм; *def* = 0, якщо дефектність відсутня. Наведемо деякі особливі види дефектності дієслівної парадигми в російській мові: відсутність синтетичних форм майбутнього часу у дієслів недоконаного виду, за винятком дієслова *быть*; відсутність форм теперішнього часу, а також дієприкметників активного та пасивного станів теперішнього часу у дієслів доконаного виду; відсутність багатьох форм безособових дієслів тощо.

Парадигматичний тип, що характеризується шістьма граматичними формами, які визначаються категорією відмінка, притаманний кількісним числівникам (крім числівника *один*). Такий парадигматичний тип будемо називати парадигматичним типом числівників:

$$W(t_4) \equiv W^C = \{w^C_1, w^C_2, \dots, w^C_6\} = \{k_i\}, i = 1, 2, \dots, 6. \quad (2.22)$$

Усі незмінні слова української та російської мови можуть бути віднесені до одного парадигматичного типу — вони мають єдину форму представлення у мові, а саме ту, яку задано в реєстровій частині словника. До незмінних слів належать усі прислівники, сполучники, прийменники, вигуки, частки, предикатні слова.

Всередені парадигматичних типів виділяємо парадигматичні класи.

Демо формальне визначення парадигматичного класу. Довільна лексема *x* (з урахуванням її словозмінних варіантів) може бути представлена у вигляді комбінації незмінної та змінної складових:

$$x = c(x) * f(x), \quad (2.23)$$

де $c(x)$ — частина лексеми x , яка в процесі словозміни залишається незмінною (квазіоснова), $f(x)$ — її змінна складова (квазіфлексія), $*$ — конкатенація.

Змінна та незмінна складові можуть мати як нульову довжину, так і представляти собою всю лексему. Наприклад, у парадигмах іменників із суплетивними формами множини (*человек, человека, ... люди, людей, ...*) незмінна частина дорівнює нулю, а змінна частина представлена всіма словоформами. У парадигмах незмінних слів, навпаки, нулеві дорівнює змінна частина.

Повна словозмінна парадигма $[x]$ слова x , що належить до парадигматичного типу t_i , представляється у вигляді:

$$\pi(x) = c(x) * \{f_i(x)\}, \quad (2.24)$$

де $f_i(x)$, $i = 0, 1, 2, \dots, n(t_i)$ — змінні частини слова (квазіфлексії) у відповідних граматичних формах; причому в деяких із них може існувати більше однієї словоформи. Для означення даного факту введемо параметр кратності граматичної форми $v(w_i(x))$, який задається цілим числом, рівним кількості можливих форм лексеми x у граматичній формі w_i . У загальному випадку:

$$f_i(x) = \bigcup_{l=0}^{v(w_i(x))} f_{il}, \quad (2.25)$$

$l=l(i) = 0, 1, 2, \dots$ — індекс кількості словоформ у граматичній формі w_i (залежить від номера граматичної форми i);

$f_{il}(x)$ — квазіфлексія початкової/вихідної форми, яка для іменника певного l конкретного роду відповідає словоформі називного відмінка однини, для дієслова — його інфінітиву, для прикметника — словоформі чоловічого роду називного відмінка однини тощо;

$n(t_i)$ — кількість граматичних форм у парадигматичному типі t_i .

Покладемо

$$\begin{aligned} F &= \bigcup_{x \in W} (\{f_0(x)\}, \{f_{1l}(x)\}, \dots, \{f_{n(t_i)l}(x)\}) \equiv \\ &\equiv \{f_{jl}^1, f_{jl}^2, \dots, f_{jl}^{N_i}\}, j=0, 1, \dots, n(t_i), l = l(w_j) = 0, 1, 2, \dots \end{aligned} \quad (2.26)$$

Тоді:

$$F = \bigcup_{k=1}^{N_i} [F]^k, \text{ де } [F]^k = \{f^k\} = \{f_{jl}^k, j=0, 1, \dots, n(t_i)\}. N_i = N(t_i), l = l(w_j) \quad (2.27)$$