

Таким чином помітки на листях дерева одержано в силу нульпо-силкових правил. Для односилкових числень одержувані у такий спосіб виводи виявляються ланцюгами.

Уточнюючи наші уявлення про числення, доходимо висновку, що скінченний список “дозвільних” правил становить хоча й най-важливішу, але тільки одну складову частину числення; він складає, так би мовити, “ядро” числення подібно до того, як оператор безпосередньої обробки складає “ядро” алгоритма.

Другою складовою частиною є інструкція щодо розділення елементів на *основні* та *допоміжні*. Назвемо цю інструкцію *правилом виділення основних станів*. Необхідність у такому правилі викликана тим, що нас — з тих чи інших мірквань — можуть цікавити не усі стани, а лише стани спеціального виду (називатимемо їх основними), тоді як інші стани розглядаються лише як допоміжний матеріал для одержання основних станів. Роль правила виділення основних станів для числень аналогічна ролі сигналу про одержання розв’язку для алгоритма.

Нарешті третьою складовою числення є *правило вилучення результату*, або *вихідна процедура*. Вона аналогічна відповідній процедурі для алгоритма. Ця процедура перетворює кожний основний стан на певний об’єкт. Результат застосування вихідної процедури до будь-якого допустимого основного стану називається *результатом* або *виходом* числення. Про кожний такий об’єкт будемо також говорити, що він *породжується численням*. Про множину всіх об’єктів, які породжуються численням, також будемо говорити, що вона породжується численням. Два числення вважаються еквівалентними, якщо вони породжують однакові множини виходів (результатів). Вважається, що множина, породжена численням, розташовується у деякому ансамблі *ансамблі виходів* розглядуваного числення.

Розглянемо приклад реалізації певного числення у вигляді логістичної системи, інакше — формальної аксіоматичної теорії. Нехай  $A$  — алфавіт цієї системи. Вважаємо відомими звичайні правила утворення (вони ж — правила побудови) та правила перетворення (вони ж — правила виводу), згідно з якими змінні ( $z$ ), терми ( $t$ ), формули ( $f$ ) та доводжувані формули ( $d$ ) виділяються з усіх слів над  $A$ . Схема побудови логістичного числення є такою. Стани в ньому мають вигляд  $\langle a, b \rangle$ , де  $a$  є однією з літер “ $z$ ”, “ $t$ ”, “ $f$ ”, або “ $d$ ”, а  $b$  є слово з  $A$ ; усі такі стани вкладаються у підходяще робоче середовище. Основними оголошуються стани вигляду  $\langle d, b \rangle$ . Правило вилучення результату — перехід від  $\langle d, b \rangle$  до  $b$ . Дозвільні правила одержуються очевидною модифікацією згаданих вище

правил утворення та перетворення. Так, правилу, яке стверджує, що будь-яка змінна є термом, відповідає однопосилкове правило, яке дозволяє перехід від будь-якого стану вигляду  $\langle p, b \rangle$  до стану  $\langle t, b \rangle$  з тим самим  $b$ ; правилу ствердження відповідає двопосилкове правило, що дозволяє перехід від  $\langle d, b_1 \rangle$  і  $\langle d, (b_1 \rightarrow b_2) \rangle$  до  $\langle d, b_2 \rangle$ . Аксиомам відповідають нульпосилкові правила: для будь-якої аксіоми  $b$  дозволяється “з нічого” зконструювати породжений об’єкт  $\langle d, b \rangle$ . Правило суперпозиції термів, яке дає за термами  $t_1, t_2$  новий терм  $u$ , приводять до двопосилкового правила, що дозволяє перехід від  $\langle t, t_1 \rangle$  і  $\langle t, t_2 \rangle$  до  $\langle t, u \rangle$ . Якщо серед правил перетворення було, скажімо, правило підстановки, яке за доводжуваною формулою  $f$ , змінною  $x$  і термом  $t$  дає нову доводжувану формулу  $g$  (яка представляє собою результат підстановки  $t$  в  $f$  замість всіх вільних входжень елемента  $x$ ), то до нашого числення повинно бути введене трипосилкове правило, яке забезпечує можливість переходу від станів  $\langle d, u \rangle, \langle p, x \rangle, \langle t, t \rangle$  до стану  $\langle d, g \rangle$ .

Останній приклад, засвідчує, що потреби у примусовому віднесенні породжуваних об’єктів до того чи іншого типу немає, оскільки інформацію про тип об’єкта можна “заховати” всередину стану, вибраного належним способом. Так, замість того, щоби породжувати терми і окремо (але за допомогою термів) формули, ми спільно породжували об’єкти вигляду  $\langle t, b \rangle$  та  $\langle \phi, b \rangle$ .

Деякі з правил, що породжують числення, задають ті чи інші перетворення (або вони є перетвореннями). Такими є правила виводу, правило вилучення результату. Інші правила задають властивості (або, коротше, є властивостями) — такими є правила виділення основних станів.

У першому розділі було визначено поняття лексикографічного числення (числення на лексикографічних структурах). Його базовим поняттям є поняття стану, лінгвістичний зміст якого ми тільки що розглядали. Зараз ми поєднаємо поняття семантичного стану з поняттям лексикографічного числення. Ідея цього поєднання полягає у використанні лексикографічних структур як джерела для параметризації семантичних станів. З цією метою розглянемо репрезентант семантичного стану, введений формулою (1.38), але у дещо видозміненій формі:

$$\psi(x_0^i) = \langle x_0^i; [M(x_0^i) \equiv \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i] \rangle, \quad (5.25)$$

де  $M(x_0^i) \equiv \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i$  — “внутрішня” частина словникової статті  $V(x_0^i)$  словникової статті СУМа зі структурою, визначеною у

третьому розділі. Множина визначених у такий спосіб станів утворює робоче середовище числення. Текст  $M(x_0^i) \equiv \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i$  вважатимемо адекватним репрезентантом повного семантичного стану лексеми  $x_0^i$ , що є цілком природним. Отже, текст  $M(x_0^i)$  містить інформацію, яка дозволяє виділити в ньому фрагменти, які представляють елементарні семантичні стани, які відповідають окремим граматичним та лексичним значенням, що, у свою чергу, дозволяє представити  $\psi(x_0^i)$  у формі (5.9): тобто

$$\begin{aligned} \psi(x_0^i) &= \mu_1(\psi_1) \psi_1(x_0^i) + \mu_2(\psi_2) \psi_2(x_0^i) + \dots + \mu_n(\psi_n) \psi_n(x_0^i) \equiv (5.26) \\ &\equiv \sum_k \mu_k(\psi_k) \psi_k(x_0^i) \end{aligned}$$

де  $\psi(x_0^i)$  — повний семантичний стан мовної одиниці  $x_0^i$ , як нечітка суперпозиція елементарних семантичних станів  $\psi_k(x_0^i)$ . При цьому:

$$\sum_k \mu_k(\psi_k) = M \quad (5.27)$$

— умова нормування, яка має простий лінгвістичний смисл як міра нечіткості у визначенні внеску кожного елементарного семантичного стану в повну семантику мовної одиниці  $x_0^i$ . Числа  $\mu_k(\psi_k)$  відображають ступінь нечіткості у визначенні внеску кожного елементарного семантичного стану в повну семантику лексеми; при  $M = 1$  їм можна надати ймовірнісну інтерпретацію. Твердження про нечіткість впливає з того, що уявлення про повноту семантики лексеми, набір і зміст елементарних семантичних станів представляють точку зору лексикографів-укладачів базової тлумачної лексикографічної системи (БТЛС), і ми не гарантовані, що враховано всі значення мовної одиниці  $x_0^i$  і що це зроблено у бездоганний спосіб.

Подальші семантичні структури можуть бути введені і обчислені через використання різних типів операторів  $A$ , з множини автоморфізмів лексикографічної системи БТЛС. Такими операторами, дію яких визначено на лексикографічній структурі БТЛС можуть виступати:

- відношення семонімії (синоніми, антоніми, пароніми, омоніми)
- відношення словотвору (однокореневі слова)
- тезаурусні відношення: [“рід — вид”, “частина — ціле”, “комплекс — елемент”, “причина — наслідок”]

— відношення асоціацій і аналогій (асоціатори і аналогеми).

Зі станів типу (5.26) можна виділити основні та допоміжні шляхом введення певної інструкції (або інструкцій) яка репрезентує певну лінгвістично обумовлену доцільність. Більше того, до визначення станів можна залучити і елементи метамови  $\tau$ . Тоді визначення лексикографічного числення можна розширити і узагальнити введенням нових об'єктів та станів. Наприклад, об'єктами можуть вже виступати не тільки окремі "лексеми" — елементи  $I_0^W(D)$ , але й інші елементи відповідної лексикографічної системи (наприклад, "формули тлумачення", їхні складові, елементи граматичного опису, ілюстрації, ремарки тощо). Правилами виводу тут можуть виступати алгоритми тестування елементів знакової системи, структури, правильності компонування словникових статей тощо. У наступному параграфі на описовому рівні буде запропоновано приклад лексикографічного числення, орієнтованого на автоматичну модифікацію та поповнення тлумачного словника шляхом його взаємодії із "зовнішнім середовищем", яким виступає масив текстів.

### 5.5.2. Узагальнена лінгвістична змінна

Нами вже було використано ідеї теорії нечітких множин при викладі поняття семантичних станів мовних одиниць. Мета цього пункту — показати, що нечіткість природним чином міститься в структурі лексикографічних систем і на основі розвиненої теорії лексикографічних структур можна сформулювати узагальнення поняття лінгвістичної змінної, яке може бути застосоване у будь-яких системах аналізу, розуміння, прийняття рішень тощо, де інформація представлена переважно її неструктурованою, природною формою.

Уточнимо та деталізуємо ці ідеї слідуючи за працями [10, 32], де розвинено інформаційний апарат, який використовується з метою обробки нечіткої інформації в системах прийняття рішень. В основу цього апарату покладено поняття лінгвістичної змінної, застосування якого дозволяє формалізувати і автоматизувати процес прийняття рішень у складних ситуаціях управління. В межах лінгвістичного підходу в якості значень змінних допускаються не тільки числа, але й слова і речення природної мови, а апаратом їх формалізації виступає теорія нечітких множин.

Наведемо основні визначення теорії нечітких множин і лінгвістичних змінних.

Нехай  $\epsilon$  універсальна множина  $U = \{u\}$ . Нечіткою множиною  $A$  на множині  $U$  назвемо сукупність пар:

$$A = \{ \langle \mu_A(u), u \rangle \}, \quad (5.28)$$

де  $\mu_A : U \rightarrow [0,1]$  — відображення множини  $U$  в одиничний відрізок  $[0,1]$ , яке називається функцією приналежності. Будемо також вживати позначення:

$$A = \bigcup_{u \in U} \mu_A(u) / u = \bigcup_{u \in U} \mu_u / u$$

Змінна  $u$  називається базовою. Можлива інтерпретація функції приналежності:  $\mu_A(u)$  є суб'єктивна міра того, наскільки елемент  $u \in U$  відповідає тому поняттю або уявленню, смисл якого формалізується нечіткою множиною  $A$ . Носієм нечіткої множини  $A$  називається множина:

$$S_A = \{ u \in U : \mu_A(u) > 0 \} \quad (5.29)$$

Нехай  $\epsilon$  дві універсальні множини  $U = \{u\}$  та  $V = \{v\}$ . Нечітким бінарним відношенням  $R$  на множині  $U \times V$  називається сукупність пар:

$$R = \bigcup_{(u,v) \in U \times V} \mu_R(u,v) / (u,v) \quad (4.28)$$

де  $\mu_R(u,v) : U \times V \rightarrow [0,1]$  — функція приналежності нечіткого відношення  $R$ , яка має той самий смисл, що і  $\mu_A(u)$ . Узагальнення на  $n$ -вимірний випадок очевидне.

Нечітка змінна визначається кортежем  $\langle A, U, \tilde{A} \rangle$ , де:

$A$  — назва нечіткої змінної;

$U$  — універсальна множина, область визначення нечіткої змінної;

$\tilde{A} = \bigcup_{u \in U} \mu_u / u$  — нечітка множина на  $U$ , яка описує обмеження

на можливі числові значення нечіткої змінної  $A$ .

Лінгвістична змінна визначається кортежем:

$$\langle N, T, U, G, M \rangle, \quad (5.30)$$

де  $N$  — назва лінгвістичної змінної;  $T$  — множина її значень або термів, які, у свою чергу, відіграють роль назв нечітких змінних з областю визначення  $U$  (базова терм-множина лінгвістичної змінної);  $G$  — “синтаксична” процедура, яка описує процес утворення з множини  $T$  нових, змістовних значень лінгвістичної змінної — позначимо через  $G(T)$  результат застосування  $G$  до  $T$ ; множину  $T^* = T \cup G(T)$  назвемо розширеною терм-множиною лінгвістичної

змінної; через  $M$  позначимо “семантичну” процедуру, яка дозволяє приписати кожному новому значенню з  $T^*$  певну семантику шляхом утворення відповідної нечіткої множини.

У залежності від характеру множини  $U$  лінгвістичні змінні поділяються на числові та нечислові. Числовою нечіткою змінною є така, для якої  $U \in R^1$  і яка має вимірювану базову змінну. Нечіткі змінні, які відповідають значенням числової лінгвістичної змінної будемо називати нечіткими числами.

Нечислові лінгвістичні змінні мають області визначення  $U$  нечислові об'єкти, зокрема, слова та інші конструкції природної мови.

Структурна теорія лексикографічних систем надає засоби для узагальнення поняття лінгвістичної змінної і введення поняття узагальненої лінгвістичної системи, основою чого служить процедура рекурсивної редукції, лексикографічна структура  $L$ -системи і поняття семантичного стану мовних одиниць.

Аналізуючи будову елементарних семантичних станів у лексикографічній системі БТЛС, доходимо висновку, що їх нижні поверхні закінчуються ілюстраціями (“мікроконтекстами”), які відіграють роль термінальних елементів інтерпретаційної частини. Отже загальне представлення елементарного семантичного стану в лексикографічній системі може бути подано формулою:

$$\psi_{MN}^Q(x) = x \rightarrow C_M^Q \rightarrow J_{MN}^Q \quad (5.31)$$

або, детальніше:

$$\psi_{MN}^Q(x) = x \rightarrow C_M^Q \begin{matrix} \nearrow J_{M1}^Q \\ \rightarrow J_{M2}^Q \\ \vdots \\ \searrow J_{Mn(N)}^Q \end{matrix} \quad (5.32)$$

де  $Q, M, N$  — групові індекси, які специфікують типи шляхів, що фігурують у визначенні структурних графів відповідних семантичних станів.

Ми вже зауважували, що у словниках, як ілюстрації, можуть використовуватися не тільки текстові сегменти, але й “фрагменти дійсності” у вигляді малюнків, відео-, аудіо-, числових даних тощо. Вважатимемо що інтерпретація нижніх (термінальних) елементів, як ілюстрацій значень відповідних елементарних інформаційних

одиниць, залишається справедливою і в лексикографічних системах загального положення. Зрозуміло, що внесок різних ілюстрацій в інтерпретацію відповідної елементарної інформаційної одиниці (тлумачення значення лексеми) не однаковий, хоча універсальних методів оцінки "вагомості" такого внеску, взагалі кажучи, немає.

Формулюючи визначення (5.31)—(5.32), ми неявно припустили, що лексикографічна структура  $\rho$  має ієрархічну будову і представляється дерев'яним графом. У загальному випадку вважаємо, що значення  $C_{(M)}^{(Q)}$  мають спільну множину терміналів  $J_{(M)\{N\}}^{(Q)}$  для всієї допустимої множини значень індексів  $\{^{(Q)}_{(M)\{N\}}\}$ . Для акуратності слід було б зробити певні уточнення стосовно належності тих чи інших  $J$  до тих чи інших  $C$ ; ввести додаткову специфікацію  $J$  за типами (і, можливо, форматами) даних, але за браком місця ми не будемо займатися цим тут, тим більше, що для цілей узагальнення поняття лінгвістичної змінної наведених визначень вже досить.

Перейдемо до формулювання поняття узагальненої лінгвістичної змінної. Механізм його досить простий і зводиться до узагальнення елементів, що фігурують у визначенні (5.30), шляхом витлумачення їх як елементів структури певних лексикографічних систем. Іншими словами, вважатимемо, що величини, представлені у формулі (5.30), визначаються згідно зі структурною теорією лексикографічних систем. Такий підхід оправдується тією обставиною, що для традиційного визначення лінгвістичної змінної (5.30) легко побудувати елементарну лексикографічну систему, структуротвірні елементи якої збігаються з наведеними у цьому визначенні елементами — пропонуємо зробити це читачеві як самостійну справу.

В результаті описаного процесу одержуємо нові, узагальнені визначення елементів лінгвістичних змінних, а саме.

(1). Множиною назв  $N$  узагальненої лінгвістичної змінної вважатимемо множину елементарних інформаційних одиниць системи  $D$  (точніше — їх назв) відносно лексикографічного ефекту  $Q$ :  $N = \{x, x \in I_0^Q(D)\}$ . Оскільки  $x \in I_0^Q(D)$  ідентифікують елементи відповідної ELS:  $x \Leftrightarrow V(x) \in V(I_0^Q(D))$ , вважатимемо що відповідність встановленою у визначенні узагальненої лінгвістичної змінної і ототожнюватимемо при необхідності множину  $N$  назв узагальнених лінгвістичних змінних з множиною пар  $\{x, V(x)\}$ . Для назв узагальнених лінгвістичних змінних встановлюємо інтерпретацію як членів (словникових статей) відповідної лексикографічної системи  $ELS[I_0^Q(D)] =$

$$= \{ I_0^Q(D); V(I_0^Q(D)) \equiv (\wedge I_0^Q(D); P(I_0^Q(D)); H; A; \lambda; \rho; \Sigma) \}.$$

(2). Множиною термів  $T(x)$  узагальненої лінгвістичної змінної  $x \in I_0^Q(D)$  вважатимемо множину її елементарних семантичних станів  $\psi^{(Q)}_{(M)\{N\}}(x)$ , які визначено формулами (5.26) — (5.27) із урахуванням уточнень (5.31) — (5.32).

(3). Базовою терм-множиною  $U(x)$  множини  $T(x) = \psi^{(Q)}_{(M)\{N\}}(x)$  узагальненої лінгвістичної змінної  $x \in I_0^Q(D)$  вважатимемо множину ілюстрацій  $J^{(Q)}_{(M)\{k\}}$ ,  $k = 1, 2, \dots, n(N)$ , — термінальних елементів системи семантичних станів  $\psi^{(Q)}_{(M)\{N\}}(x)$ . З контексту визначення величин  $J^{(Q)}_{(M)\{k\}}$  зрозуміло, що вони можуть набувати як числових, так і нечислових значень, тобто мають комбінований характер.

(4). Функцією приналежності базової терм-множини  $U(x)$  є відображення:

$$\mu^J(x): J^{(Q)}_{(M)\{k\}} \rightarrow [0,1],$$

введення якого перетворює базову терм-множину  $U(x)$ , а з нею і всю узагальнену лінгвістичну змінну на нечіткий об'єкт.

(5). Для узагальненої лінгвістичної змінної доцільно визначити певну додаткову функцію, а саме — функцію приналежності множини  $T(x)$

$$\mu^\psi(x): \psi^{(Q)}_{(M)\{N\}}(x) \rightarrow [0,1],$$

введення якої перетворює множину термів  $T(x)$  на нечіткий об'єкт.

(6). Множина процедур розширення системи  $I_0^Q(D)$ :  $W = \{w_i, i = 1, 2, \dots\} = \{A; \pi; \dots\}$ , де  $A$  — множина автоморфізмів  $ELS$ ;  $\pi$  — породжуюча функція —  $\pi: I_0^Q(D) \rightarrow \hat{I}^Q(D)$ .

$$w_i: I_0^Q(D) \rightarrow \hat{I}^Q(D); I_0^Q(D) \subseteq I_0^Q(D) \subseteq \hat{I}^Q(D).$$

(7). Множина  $M$  — семантична процедура — надання змісту лінгвістичних змінних елементам розширеної системи  $\hat{I}^Q(D)$ .

Визначення 2. Твердження (1)—(7) визначають узагальнену лінгвістичну змінну.

На основі поняття узагальненої лінгвістичної змінної природно формулюється поняття лінгвістичної системи. Робиться це у такий спосіб.

Визначення 3. Лінгвістична система визначається кортежем:

$$\langle N; V; T; U; \mu^J; \mu^\psi; W; M \rangle, \quad (5.33)$$

елементи якого інтерпретуються так:

$N = \{x, x \in I_0^Q(D)\}$  — множина назв узагальнених лінгвістичних

змінних, роль якої відіграє множина елементарних інформаційних одиниць певної (базової) лексикографічної системи;

$V \equiv \text{ELS}[L] = \{ I_0^Q(D); V(I_0^Q(D)) \equiv (\Lambda(I_0^Q(D)); P(I_0^Q(D))) \}; H; A; \lambda; \rho; \Sigma \}$  — базова елементарна лексикографічна система;

$T = \cup T(x)$  — сукупність множин термів узагальнених лінгвістичних змінних,  $x \in I_0^Q(D)$ ;

$U = \bigcup_{x \in I_0^Q(D)} (x)$  — сукупність базових терм-множин для  $T$ ;

$\mu^J = \{ \mu^J(x), \forall x \in I_0^Q(D) \}$ ;

$\mu^W = \{ \mu^W(x), \forall x \in I_0^Q(D) \}$ ;

$W$  — множина процедур розширення системи  $I_0^Q(D)$

$M$  — множина процедур надання змісту лінгвістичних змінних елементам розширеної системи  $\hat{I}^Q(D)$ .

Лінгвістична система являє собою певний тип нечіткої інформаційної системи, оскільки вона справді є інформаційною системою (як лексикографічна система, що складає її субстрат), в якій розвивається цілий ряд нечітких відношень.

Наведемо приклади узагальнених лінгвістичних змінних та узагальнених лінгвістичних систем.

У лексикографічній системі Словника української мови, структуру якої ми тут використовуємо у тій формі, як її побудовано у п'ятому розділі книги<sup>15</sup>, індукується структура лінгвістичної системи в такий спосіб.

Назвами узагальнених лінгвістичних змінних тут виступають реєстрові одиниці СУМа, які складають множину  $I_0(U)$ .

Для кожного  $x \in I_0(U)$  відповідна терм-множина  $T(x)$  визначається формулами:

$$\psi^J_i = (x) \rightarrow C_i, i = 1, 2, \dots, n(x) \quad (5.34)$$

$$\psi^V_{ik} = (x) \rightarrow C_i \rightarrow V \rightarrow V_{ik}, k = 1, 2, \dots, n(i)$$

$$\psi^F_{ip} = (x) \rightarrow C_i \rightarrow FC \rightarrow FC_{ip}, p = 1, 2, \dots, n(k)$$

$$\psi^{FV}_{ipr} = (x) \rightarrow C_i \rightarrow FC \rightarrow FC_{ip} \rightarrow FC_{ipr}, r = 1, 2, \dots, n(p)$$

які одержуються з формул (5.32)—(5.32) редукцією останніх за відповідними елементами  $J$ .

Базовою терм-множиною  $U(x)$  терм-множини  $T(x)$  служить об'єднання множин відповідних ілюстрацій:

$$\{ J_{ij} \} \cup \{ J^V_{ikl} \} \cup \{ J^{FC}_{ipq} \} \cup \{ J^{FCV}_{iprs} \} \quad (5.35)$$

<sup>15</sup> В.А.Широков. Феноменологія лексикографічних систем

Визначення  $W$  — множини процедур розширення системи  $I_0(U)$  для СУМу може включати такі елементи:

- відношення парадигматизації;
- тезаурус над  $I_0(U)$  [відношення: “рід — вид”, “частина — ціле”, “комплекс — елемент”, “причина — наслідок” і т. ін.];
- відношення синонімії;
- керування та узгодження;
- еквіваленти слова;
- фразеологізми і т. ін.

Функція приналежності  $\mu^W$  визначається за формулами (5.26) — (5.27).

Процедури визначення конкретних значень  $\mu^W$  можуть використовувати механізми експертних оцінок при укладанні (формуванні) лінгвістичної системи або інші механізми, але треба враховувати, що значення  $\mu^W$  можуть мінятися в процесі семантичного аналізу в залежності від цілого ряду чинників: введення додаткової інформації, ідентифікації з аналізованого тексту певних елементарних семантичних структур, що одразу веде до редукції розподілу (5.26) — (5.27) тощо. Визначення і введення до семантичного аналізатора зазначених механізмів надає всій системі необхідної гнучкості, чим зближує її поведінку з поведінкою людських аналітичних систем. За допомогою викладеного формалізму легко переносяться і узагальнюються поняття, твердження і результати, що стосуються лінгвістичних критеріїв, нечітких висловлювань, лінгвістичних лотерей, інформаційних гранул тощо.

# БАГАТОМОВНІ ЛЕКСИКОГРАФІЧНІ СЕРЕДОВИЩА

## 6.1. Представлення перекладних еквівалентів у багатомовному перекладному словнику (БПС)

### 6.1.1. Принципи побудови БПС у концепції семантичних станів

Експліцитними представниками семантичних станів одиниць перекладного словника слугують їх конкретні значення, зафіксовані у словнику. У БПС згідно з роботою<sup>1</sup> словосполучення не представлено, оскільки перекладні словники словосполучень (ПСС) укладаються для кожної пари мов окремо з орієнтованим напрямом перекладу, і при автоматичному перекладі перевірка за ПСС передує пошукові перекладних еквівалентів (ПЕ) за БПС. У системі МП БПС та ПСС інтегруються в єдину Л-систему підтримки МП.

ПЕ у відповідних зонах БПС представлено синонімічними групами слів у смислі формули (5.10), яка може реалізовуватися у доволі широкому класі лексичних контекстів даної мови. Отже семантичний стан синсета, як репрезентанта перекладного еквівалента, визначаються такими характеристиками: а) близькість значень окремих елементів групи і б) взаємозамінність у тексті, яка не веде до істотної зміни його смислу.

Практичне формування синсетів є роботою доволі непростою, зважаючи на відзначену вище відмінність між семантикою тлумачних словників і семантикою синонімії та великим обсягом лексикографічної інформації, яку доводиться опрацьовувати у процесі формування репрезентативної множини синсетів. З метою збільшення ефективності цієї роботи в Українському мовно-інформаційному фонді створено автоматизований інструментальний комплекс, котрий являє собою комп'ютерну лексикографічну систему, яка реалізує певну схему інтеграції електронних версій тлумачного Словника української мови (СУМ)<sup>2</sup> і Словника синонімів української мови (ССУМ)<sup>3</sup>. Виклад відповідної системотехніки буде здійснено далі,

<sup>1</sup> Т.О. Грязнухіна. "Мовознавство" № 5 2001.

<sup>2</sup> Словник української мови в 11 томах.

<sup>3</sup> Словник синонімів української мови.

а зараз перейдемо до аналізу принципів, які ми застосовуємо при поданні синонімії у конструкції БПС.

Виходячи з вимоги взаємозамінності при перекладі елементів всередині синсета, з нього, як правило, виключаються лексеми, які перебувають у родо-видових відношеннях. Наприклад, синонімічні групи, задані в ССУМ із заголовковим словом **“вітер”**: 1). {**вітер, вітровий, вітрило**}, 2). {**вітер, легіт, легковий, зефір, повітер, бриз**}, 3). {**вітер, фен**}, 4). {**вітер, суховій**}, 5). {**вітер, степовик, степняк**}, 6). {**вітер, буровій, борвій**}, 7). {**вітер, поземок, поземка**}. Після аналізу семантики цих груп за тлумачним словником, третя та четверта групи вилучаються, оскільки другі члени кожної з них позначають різновиди вітру (**фен** — *сухий і теплий, що дме з гір в долину; суховій* — *сухий і гарячий*), і отже, перебувають у відношенні рід-вид; з другої групи було вилучено **“вітер”** і **“бриз”** (*береговий морський вітер*), а ті слова, що залишилися у цій групі було поєднано значенням **“легкий вітерець”**. З п'ятої, шостої й сьомої груп вилучено лексему **“вітер”** і групам надано, відповідно, значення **“степовий вітер”**, **“сильний вітер”**, **“низовий вітер узимку, що переносить по землі сніг”**.

Вимога взаємозамінності диктує видалення з синсетів тих лексем, семантичні стани яких характеризуються дуже вузьким колом контекстів. Наприклад, з дієслівного синсета **“повертати, звертати, вернути, брати, забирати”** зі значенням за ССУМ **“змінювати напрямок свого руху, повертати”** дві останніх лексеми вилучаються, оскільки лексема **“брати”** у цьому значенні можлива лише у контекстах **“брати вліво, вправо”**, а **“забрати”** — **“забрати вліво, вправо, вгору”** (пор.: *повертати вліво, вправо, з дороги, на стежку, до лісу, в напрямку, від дуба...*). Група ССУМа {**виконувати, нести**} зі значенням **“здійснювати що-небудь, реалізувати завдання, наказ, задум тощо”** взагалі не задовольняє вимозі взаємозамінності, оскільки непрямі значення лексеми **“виконувати”**: **“нести відповідальність”**, **“нести військову службу”** у системі семантики лексеми **“нести”** взагалі фразеологізуються.

У структурі Л-системи БПС перекладним еквівалентам певного синсета присвоюється однаковий номер у всіх мовних зонах, чим забезпечується, по-перше, адекватність перекладу слів з однієї мови на іншу, і по-друге, можливість формального розмежування варіантів перекладу, а також синонімів до цих варіантів. Наприклад, серед варіантів перекладу українського прикметника **“рівний”** на російську мову не буде ПЕ **“сдержанный”**, **“уравновешенный”**, **“выдержанный”** — всі вони будуть представлені одним з цих слів (скоріше всього **“уравновешенный”**), а решта буде розглядатися

як елементи синсета із заголовковою лексемою “уравновешений”. Обраний автоматично варіант перекладу на етапі постредугування може за бажанням користувача бути замінений одним з синонімів відповідної групи.

Кількість можливих варіантів перекладу слова обмежується кількістю його семантичних станів, зафіксованих у значеннях синсетів, котрі містять слово, яке підлягає перекладу. Всередині кожного синсету доцільно встановлення порядку (бодай часткового) за ступенем віддаленості членів синсета від головного (спільного) значення. Ранжування, наприклад, може здійснюватися у такий спосіб: ранг 1 приписується слову з найбільш нейтрально вираженим значенням у синсеті. Стилістично маркованим елементам синсета (наприклад, таким що мають поміти у словнику “розм.”, “рідко” тощо надаються найвищі ранги. Морфолого-фонетичним варіантам додається спеціальний ідентифікатор; наприклад у варіанті БПС, що розробляється в Українському мовно-інформаційному фонді перед рангом такого варіанту ставиться цифра 5. За збігом другої цифри з рангом іншого слова і відбувається ідентифікація “абсолютних” синонімів. Наведемо приклад ранжування всередині синсета. Так синсет зі значенням “бути спроможним зробити щось за певний проміжок часу” набуває з урахуванням рангів такого вигляду: {встигати 1, устигати 51, вкладатися 2, укладатися 52, поспівати 3, постигати 4}.

У системі МП передбачено процедуру вибору основного варіанту перекладу з урахуванням кількості в групі синонімічних ПЕ та номерів рангів слова, що перекладається, всередині синсетів. У мові-мішені як варіанти перекладу насамперед пропонуються слова з рангом 1.

Для розв’язання проблеми зменшення числа варіантів перекладу, котрий здійснюється за підтримкою автоматичного перекладного словника — в ідеалі воно дорівнюватиме 1, до системи детермінантів семантичного стану ПЕ пропонується ввести певні параметри граматичної семантики, а також прагматичного контексту — останні покликані забезпечити в алгоритмі перекладу ефект детермінації лексичних варіантів перекладу текстовою ситуацією вхідної мови.

У згаданому варіанті БПС введено такі параметри граматичної семантики: 1) ознака лексико-граматичного класу для всіх одиниць словника; 2) для іменників — ознаки роду, істотності, числа (якщо слово має це значення тільки в однині або тільки у множині); 3) для дієслів — ознаки виду та перехідності; 4) для дієслів, іменників та прикметників — моделі керування (прийменникового

і безприменникового) із вказанням ознаки істотності/неістотності для залежної форми, якщо ця ознака спроможна бути диференційною для конкретного значення багатозначного слова. Наприклад, при перекладі в українських фразах "...бачив, як пан бив, нівечив своїх кріпаків" і "нівечили усе підряд — знаряддя, машини..." дієслова "нівечити", яке в українській зоні БПС входить до трьох синсетів<sup>4</sup>, у першому випадку за ознакою при "нічевити" Зн.1 (сильне керування залежною формою з ознакою "істот.") і ознакою у слова "кріпаків" "істот." у всіх зонах БПС буде встановлено односторонній переклад, а у другому випадку з трьох можливих варіантів перекладу буде вибрано два, які встановлюються за двома останніми синсетами.

Прагматичний контекст пропонується задавати ознакою підмови, де слово вживається найчастіше. Якщо систему МП орієнтовано на переклад текстів, індексованих за параметром "тематична група", або коли система сама має засоби автоматичного тематичного індексування, то за збігом індексу тексту та відповідної мітки біля слова, що перекладається, в одному із синсетів саме цей останній і визначатиме вибір основного варіанту перекладу. Наприклад, при перекладі українського прикметника "рівний", семантичні стани якого в українській зоні БПС зафіксовано в 7-ми синсетах без урахування вказаного параметра, при перекладі на російську мову одержуємо 7 варіантів перекладу: 1) **одинаковий** (синонім — **равный**); 2) **равный**; 3) **прямой** ("ровный"); 4) **гладкий** ("ровный"); 5) **плавный** ("ровный, гладкий"); 6) **спокойный** ("ровный"); 7) **сдержанный** ("уравновешенный, выдержанный, ровный"). Із урахуванням мітки "матем." біля слова "рівний" в одному із синсетів при перекладі текстів з такою ж міткою сім варіантів перекладу редукуються до двох: "равный" та "одинаковий".

### 6.1.2. Приклади представлення перекладних еквівалентів у зонах БПС.

Синсети, що відповідають семантичним станам українського іменника "земля": "третя від Сонця планета", "верхній шар земної кори", "те, що не є водняним простором", "територія, що становить

<sup>4</sup> 1). "Заподіювати комусь муки, знущання": {злушчатися (над+ор. 2 -ким), глумитися (над+ор. 2), нівечити (зн. 2 — кого), мордувати (зн. 2), катувати (зн.2), мучити (зн. 2)}; 2). "Робити непридатним для користування, вживання": {псувати(зн.1), спотворювати (зн.1), ушкоджувати (зн.1), пошкоджувати (зн.1), нівечити (зн.1)}; 3). "Показувати щонебудь у неправильному вигляді": {викривляти (зн.1), спотворювати (зн.1), перекручувати(зн.1), нівечити(зн.1)}.

адміністративно-політичну єдність”, “місцевість, область, район, що має певні кліматичні і природні особливості”, в українській, російській та англійській зонах БПС набувають такого вигляду:

№	СИНСЕТИ В МОВАХ:		
	Українській	Російській	Англійській
47	Земля ( <i>тільки одн.</i> )	Земля (только ед.ч.)	world earth universe
49	грунт земля	почва земля	soil ground
50	суходіл суша земля материк	суша земля материк	land
51	держава країна край земля (только ед.ч.) сторона	государство держава страна земля (только ед.ч.) край сторона	state country land nation
117	край місцевість сторона країна	страна край сторона местность	place land country locality surroundings

З використанням наведених лексикографічних структур ось як виглядатиме варіант перекладу з української на російську слів “країна” та “земля” у реченні “Демократичних країн збільшилося на Землі”:

**Країна:** 1) государство (синоніми: держава страна земля (*только ед.ч.*) край сторона); 2) страна (синоніми: край сторона местность);

**Земля:** 1) Земля (*только ед.ч.*); 2) государство (синоніми: держава страна земля (*только ед.ч.*) край сторона) 3) суша (синоніми: земля материк) 4) почва (синонім: земля).

Звідси впливають варіанти перекладу, яким віддається перевага:

країна → государство

та

Земля → Земля.

## 6.2. Інтегрована система синонімічної та тлумачної Л-систем

### 6.2.1. Принципи побудови WordNet

У зв'язку з роллю, яку відіграє відношення синонімії у процесах перекладу, виникає необхідність мати репрезентативний, добре структурований та узгоджений із тлумачним комп'ютерний словник синонімів. Словники синонімів є досить традиційними різновидами лексикографічних праць, проте останнім часом і у цій галузі є певні зрушення. Йдеться про так звану систему WordNet, яку

підтримує всесвітня асоціація Global WordNet. Робота над цією системою розпочалася у 1985 році в лабораторії когнітології Принстонського університету. Спочатку систему WordNet було створено для англійської мови, але в середині 90-х років був заснований проєкт EuroWordNet, що став розширенням WordNet та дав можливість на базі його принципів створити аналогічні системи для інших мов. Структура WordNet побудована у такий спосіб, що набір загальних понять у ній не залежить від особливостей конкретних мов, тому жодна мова не має пріоритету.

На сьогодні Global WordNet представлена багатьма організаціями (більшість з них також входять до EuroWordNet), які створили WordNet-системи для французької, німецької, італійської, португальської, іспанської, шведської, голландської, баскської, чеської та інших світових мов. Також у стані розробки знаходяться проєкти BalkanNet (болгарська, грецька, молдавська, румунська, словенська, турецька мови), IndianNet (тамільська, мова каннада), NordicNet (ісландська мова), аналогічні проєкти започатковано для російської, латвійської, іврити та інших мов, які готуються вступити до Global WordNet. Для України, яка заявила про свій європейський вибір, створення національної системи WordNet та входження до Global WordNet та EuroWordNet є актуальною проблемою.

WordNet представляє собою діалогову лексичну базу даних системи іменників, дієслів, прикметників та прислівників, яка базується на психолінгвістичних теоріях людської пам'яті. Принцип упорядкування лексико-семантичної інформації у цій системі на є алфавітним, оскільки звичайний алфавітний принцип впорядкування лексично зібраної інформації розводить слова зі спільною, релевантною семантикою. Таким чином, організація даних, яку дозволяють традиційні лексикографічні праці, виявляється з огляду на значення випадковим списком слів, близьких за написанням. Пошук будь-якої лінгвістичної інформації за алфавітом втомлює і забирає багато часу, і очевидно — зумовлює неточності та недогляди. Сучасні комп'ютерні технології пропонують альтернативу, яка дозволяє лексикографам, а далі і користувачам швидко знаходити необхідну семантичну інформацію. Вхід до традиційних словників зазвичай може бути здійснений лише через реєстрове слово. Можливості ж комп'ютерних систем дозволяють, значно заощаджуючи час, коректно виявляти будь-які дані, що містяться у лінійному тексті словників. Передумовою для цього повинна стати конвертація тексту традиційних словників до відповідної форми — лексикографічної бази даних, яка забезпечує доступ до будь-яких даних, що містять-

ся у ній. Отже, WordNet став спробою ефективного поєднання традиційної лексикографічної інформації та сучасної швидкодії комп'ютерних технологій.

Основою створення системи WordNet для англійської мови послужив Оксфордський словник англійської мови Меррея, який вирішує проблеми щодо значення та вживання слів, однак у ньому акцентовано увагу на історичному аспекті й залишено поза увагою питання синхронної організації лексичних знань. Сьогодні можна уникнути таких недоліків і таке завдання було розв'язано саме застосуванням психолінгвістичних принципів, коли у 1985 році група фізіологів та лінгвістів Принстонського університету зайнялася створенням лексичної бази даних, головна ідея якої полягала в тому, щоб упровадити поняттєвий, а не алфавітний принцип її упорядкування. Результатом цього й стала поява WordNet, який можна вважати словником, що базується на психолінгвістичних принципах.

Зауважимо, що спосіб застосування психолінгвістичної теорії до лексикографічної справи був усвідомлений не одразу стало, оскільки більшість психолінгвістичних досліджень до того часу поширювалася на невеликі частини лексику (переважно на певні групи іменників). Таким чином, одним із первинних завдань WordNet стало поширення психолінгвістичних гіпотез на повний діапазон лексику.

За ідеологією WordNet всю множину слів розподілено за п'ятьма лексико-граматичними класами, а саме такими: іменники, дієслова, прикметники, прислівники та так звані функціональні слова (останні фактично зараз не представлено в системі). Причому спосіб структурування кожної з названих категорій є індивідуальним. Так, виявилось, що іменники організовано у лексичній пам'яті як тематичні ієрархії, дієслова організовано як різновид імплікативних відношень, а прикметники і прислівники — як N-вимірні гіперпростори. Кожна із зазначених структур відображає різний шлях й спосіб категоризації мовного матеріалу. Спроби накладання єдиного принципу організації на всі синтаксичні категорії не є релевантними, оскільки вони спотворюють психологічну складність лексичного знання. Тому підсистеми мовної системи будуються в межах певної частини мови, що накладає свою специфіку на організацію лексикографічних структур. Частиномовну стратифікацію мовної системи підтверджує й експериментальне вивчення словесних асоціацій.

Специфікою WordNet є його спроба організувати лексичну інформацію перш за все в термінах значення слова, а не його форми.

