Т. П. ЛЮБЧЕНКО, І. В. ШЕВЧЕНКО, К. В. ШИРОКОВ

ГРАМАТИЧНІ ЛЕКСИКОГРАФІЧНІ СИСТЕМИ З ФЛЕКТИВНИМ І АГЛЮТИНАТИВНИМ КОМПОНЕНТОМ СЛОВОЗМІНИ

У статті подано концептуальні моделі словозмінної системи мов із флективним компонентом словозміни, мов із флективним та аналітичним компонентами, а також мов з аглютинативним компонентом. Описано технологію побудови граматичних лексикографічних систем, їхню структуру та функції.

Ключові слова: граматична лексикографічна система, граматичний словник, парадигматичний тип, граматичний клас, парадигматичний клас, дефектність словозмінної парадигми, варіативність словозмінної парадигми.

Граматичні лексикографічні системи (ГЛС) спрямовані на використання їх у процесах автоматизованого опрацювання мовної інформації на етапах визначення граматичного статусу слів, їхньої вихідної форми й синтезування словозмінних форм. Особливої актуальності це завдання набуває для мов із флективним компонентом у словозміні, зокрема для слов'янських, а також і для аглютинативних, які мають складну словозмінну систему.

В Українському мовно-інформаційному фонді НАН України (УМІФ НАНУ) створено лексикографічні системи граматичного типу для української, російської, німецької, турецької, англійської мов, а також розробляються такі системи й для іспанської, французької, польської, білоруської мов.

Концептуальну основу розробок становить теорія лексикографічних систем [1], що дозволяє з єдиних позицій будувати граматичні лексикографічні системи для різних мов (не тільки флективних, а й флективних з елементами аналітизму, аглютинації). Концептуальні моделі, які покладено в основу створюваних систем, ураховують широкий спектр параметрів, що забезпечує гнучкість цих систем і уможливлює виконання основних їхніх функцій, спрямованих на автоматичне опрацювання природної мови. Це дозволяє також здійснювати глибокі дослідження словозмінних систем флективних мов, зокрема щодо таких параметрів словозмінної парадигми, як її дефектність (неповнота) і варіативність, а також забезпечує визначення повного системного опису широко представленого в цих мовах явища граматичної омонімії. Зокрема, отримані дані дослідження щодо неповноти й варіативності можуть бути використані в типологічних дослідженнях словозмінних систем відповідних флективних мов.

Для підтримки граматичних лексикографічних систем в УМІФ НАНУ створено інструментальні комплекси, що дозволяють у зручній формі виконувати

[1] *Широков В. А.* Інформаційна теорія лексикографічних систем.— К., 1998.— 331 с.

лексикографічну роботу з супроводу лексикографічних баз даних відповідних мов. Для української, російської мов такі інструментальні комплекси працюють у режимі віртуальної лексикографічної лабораторії, що уможливлює ефективну взаємодію між користувачами та компонентами системи.

Мова з флективним компонентом. Згідно з теорією лексикографічних систем, будова словозмінної лексикографічної системи (ЛС) для мови з флективним компонентом може бути представлена діаграмою (рис. 1):

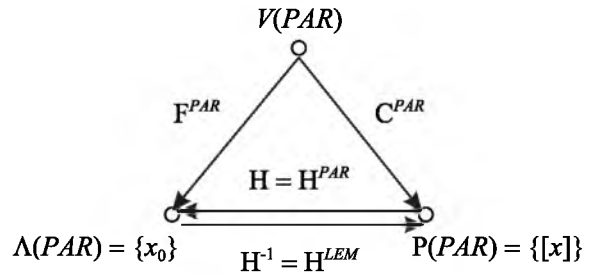


Рис. 1. Структура морфологічної (парадигматичної) ЛС.

У наведеній схемі елементи мають таку інтерпретацію: $V(PAR)$ — множина словникових статей морфологічної (парадигматичної) ЛС; $\Lambda(PAR)$ — множина лівих частин словникових статей: $\Lambda(PAR) = F^{PAR}V(PAR) = \{x_0\}$, де $\{x_0\}$ — множина слів у вихідній формі; $P(PAR)$ — множина правих частин словникових статей: $P(PAR) = C^{PAR}V(PAR) = \{[x]\}$, де $\{[x]\}$ — множина всіх словозмінних парадигм [2]; $H = H^{PAR}$ — оператор парадигматизації, який визначає відповідність між вихідною формою слова x_0 і множиною його словозмінних форм $[x]$: $H^{PAR}x_0 = [x]$; $H^{-1} = H^{LEM}$ — оператор лематизації: $H^{LEM}\chi(x) = x_0$, де $\chi(x)$ — довільний елемент парадигми $[x]$.

Власне, завдання побудови ЛС граматичного типу й полягає в побудові оператора парадигматизації, що в свою чергу вимагає формалізації словозмінної системи природної мови (тобто побудови формальної моделі словозміни).

Моделювання словозміни мови флективного типу починається з визначення та формалізації лінгвістичних критеріїв, відповідно до яких уся множина слів мови розбивається на певні підмножини, що взаємно не перетинаються, усередині кожної з яких словозміна підпорядковується єдиному правилу. Підмножини слів з такими властивостями називаються словозмінними (парадигматичними) класами.

Уведемо позначення: L — певна мова із флективним словозмінним компонентом, W^L — клас слів мови L. Моделювання розбиття сукупності слів певної фіксованої мови на парадигматичні класи здійснюється в кілька етапів. На першому етапі лексикон розбивається на класи слів за типом словозмінної парадигми. Такі класи слів у подальшому називатимемо парадигматичними типами (ПТ). Кожний парадигматичний тип характеризується набором ознак, що їх виражають граматичні категорії, граматичні значення і форми [3]. Позначимо парадигматичні типи символом T_i, де $i = 1, 2, \ldots, N$; N — кількість парадигматичних типів; $W^L(T_i)$ — множина слів мови L, що належить типу T_i (має тип T_i), а

[2] Словоформи, що входять до складу словозмінної парадигми, називаємо парадигматичними формами.

[3] *Виноградов В. В.* Русский язык // Грамматическое учение о слове.— М., 1986.— 640 с.; Лингвистический энциклопедический словарь / Гл. ред. В. Н. Ярцева.— М., 1990.— 685 с.

$K = \{K_1, K_2, .., K_n\}$ — множина граматичних категорій мови. Отже, формальне визначення парадигматичного типу T можна записати так: $<x \mid T \mid K> = \{<x, K_1> \delta(x, K_1); .., <x, K_n> \delta(x, K_n)\}$, де функція $\delta(x, K_i)$ дорівнює 1, якщо x має властивість K_i, де K_i — словозмінна граматична категорія; дорівнює 0, якщо x не має властивості K_i, або x має властивість K_i, але K_i не є словозмінною граматичною категорією.

Відповідно до словозмінних категорій, що визначають словозмінну парадигму конкретних слів, парадигматичні типи вводяться таким чином, що

$$W^L = \bigcup_{i=1}^{N} W^L(T_i) \text{ і } W^L(T_i) \cap W^L(T_j) = \varnothing, i \neq j,$$

тобто клас слів мови L складається з підкласів $W^L(T_i)$, які взаємно не перетинаються і містять слова, що мають той самий парадигматичний тип.

Детальний опис парадигматичних типів кожної з розглядуваних нами мов можна знайти у відповідних публікаціях [4]. Для російської мови, наприклад, характерними є такі парадигматичні типи: субстантивний, ад'єктивний, дієслівний, кількісних числівників і так званий нульовий парадигматичний тип (незмінювані слова мови).

За ознакою належності до певної частини мови (лексико-граматичного класу) і за додатковими ознаками, які є класифікувальними в її межах, множина слів W^L розбивається на підмножини, що їх ми називаємо *граматичними класами*. У подальшому викладі позначатимемо їх $P_j, j = 1, 2, ..., p$, де p — кількість граматичних класів; $W^L(P_j)$ — слова мови L, які належать до граматичного класу P_j.

Розбиття на граматичні класи здійснюється таким чином, що

$$W^L = \bigcup_{j=1}^{p} W^L(P_j). \quad (1)$$

На цьому етапі розгляду вважатимемо омонімію знятою, а омоніми промаркованими. Тоді взаємний перетин слів з різних граматичних класів є пустою множиною при $W^L(P_i) \cap W^L(P_j) = \varnothing$ при $i \neq j$, $j = 1, 2, ..., p$.

Між парадигматичними типами і граматичними класами існує зв'язок, який реалізується відношенням вкладення: $P_j \subseteq T_i$, де P_j, $j = 1, 2, ..., p$ — граматичні класи, в яких словозміна має парадигматичний тип T_i, p — кількість відповідних граматичних класів.

Принцип розподілу лексики флективної мови на парадигматичні типи, граматичні і парадигматичні класи подано на рис. 2.

Певний парадигматичний тип може бути властивий одному або кільком граматичним класам, усередині кожного з яких виокремлюються парадигматичні класи. Як видно зі схеми, відповідно до принципу розподілу

[4] *Любченко Т. П.* Моделирование морфологии естественного флективного языка // Бионика интеллекта.— 2008.— № 1.— С. 52–64; *Lyubchenko T.* Modelling of the Digital Grammar Dictionary of Russian // Organization and Developmemt of Digital Lexical Resources.— K., 2009. — P. 73–84; *Шевченко І. В.* Параметризація як основа граматичної ідентифікації словникових одиниць української мови // Прикладна лінгвістика та лінгвістичні технології : Зб. наук. праць.— К., 2008.— С. 393–404; *Shevchenko I.* Towards Creation of the Polish Grammatical Dictionary // Organization and Developmemt of Digital Lexical Resources.— K., 2009.— P. 61–65; *Shevchenko I., Kotsyba N., Kurshuk K.* Towards the Creation of a Belarusian Grammatical Dictionary // Explorations Across Languages and Corpora. — Frankfurt am Main et al. — 2011. — Vol. 24.— P. 547–562; *Шевченко І. В.* Моделі та алгоритмічно-програмне забезпечення лексикографічних систем : Дис. ... канд. техн. наук. — К., 2000.— 167 с.; *Широков К. В.* Именна словозміна в сучасній турецькій мові.— К., 2009.— 318 с.; *Любченко Т. П.* Лексикографічні системи граматичного типу та їх застосування в засобах автоматизованого опрацювання природної мови : Дис. … канд. техн. наук.— К., 2011.— 294 с.

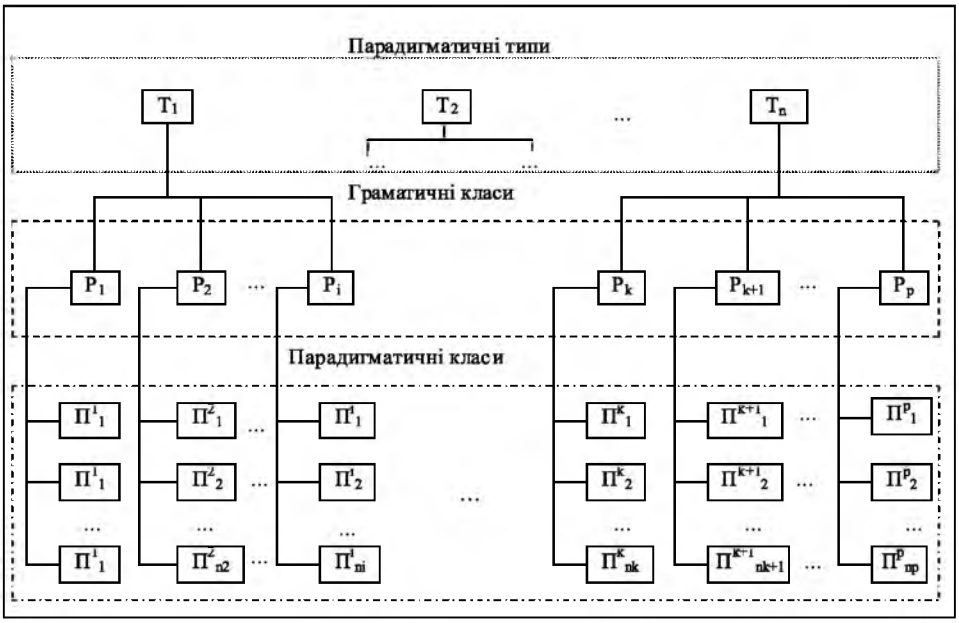


Рис. 2. Загальний принцип розподілу лексики флективної мови

$$W^L(T_i) = \bigcup_{j=1}^{p_i} W^L(P_j), \tag{2}$$

де $P_j \subseteq T_i$, p_i — кількість граматичних класів, у яких словозміна має парадигматичний тип T_i; а

$$W^L(P_j) = \bigcup_{k=1}^{n_j} W^L(\Pi_k), \tag{3}$$

де $\Pi_k \subseteq P_j \subseteq T_i$, n_j — кількість парадигматичних класів граматичного класу P_j, у яких словозміна має парадигматичний тип T_i.

Парадигматичний клас формально визначаємо таким чином: для мови із флективним компонентом словозміни довільну лексему x (з урахуванням її словозмінних варіантів) можна розглядати як таку, що складається з незмінюваної та змінюваної частин:

$$x = c(x) * f(x), \tag{4}$$

де c(x) — частина лексеми x, яка в усіх словозмінних формах залишається незмінюваною (квазіоснова), $f(x)$ — її змінюваний складник (квазіфлексія), символом «*» позначено операцію конкатенації. Змінюваний та незмінюваний складники лексеми можуть мати як нульову довжину, так і повністю збігатися з лексемою. Наприклад, у російській мові в парадигмах іменників із суплетивними формами множини (*человек*, *человека*, ..., *люди*, *людей*, ...) квазіоснова має нульову довжину, а квазіфлексії представлені всіма словоформами. У парадигмах незмінюваних слів, навпаки, нульову довжину має квазіфлексія.

Словозмінна парадигма лексеми x репрезентована як

$$\pi(x) = c(x) * \{f_j(x)\}, \tag{5}$$

де $f_j(x), j = 0, 1, 2, ..., n(T_i)$ — змінювані частини слова у відповідних граматичних формах, причому в деяких з них може існувати більше однієї словоформи, тобто:

$$f_j(x) = \{f^l_j\}, \tag{6}$$

де $l = l(j) = 0, 1, 2, ..., \nu$ — кратність граматичної форми.

Для розбиття множини слів певного граматичного класу на парадигматичні класи будуються відношення парадигматизації π_i, що визначаються так:

$$\forall\, x^1, x^2 \in P_i \;\; x^1 \pi_i x^2 : x^1 = c(x^1) * f^k, x^2 = c(x^2) * f^k, f^k \in [F]^k, \qquad (7)$$

де $[F]^k$ — набір квазіфлексій, який характерний для певної групи слів, що мають у відповідних граматичних значеннях певного парадигматичного типу однакові змінювані складники. Відношення парадигматизації є відношенням еквівалентності, оскільки воно має властивості рефлексивності, симетричності й транзитивності. Фактор-множина P_i/π_i становить множину парадигматичних класів $\{\Pi_j\}$ граматичного класу P_i (парадигматичного типу T_i):

$$P_i = \bigcup_{j=1}^{n} \Pi_j.$$

До одного парадигматичного класу належать слова, що мають однакові набори квазіфлексій для відповідних граматичних форм і відрізняються один від одного лише незмінюваним складником $c(x)$. Слова з одного класу еквівалентності, який визначено в такий спосіб, мають відповідно й однакові правила словозміни.

Для автоматичного отримання повної парадигми за її початковою формою будується оператор парадигматизації:

$$\mathrm{H} : x_0 \to [x] = c(x) * \{f_0(x), f_1(x), ..., f_n(x)\} \equiv \{c(x) * f_0(x), c(x), ..., c(x) * f_n(x)\}, \qquad (8)$$

дія якого визначається відношенням π (x^1, x^2).

Оператор парадигматизації для кожного парадигматичного класу визначається незалежно:

$$\forall x \in W(\Pi_k) \subseteq W(P_j) \subseteq W(T_i), \mathrm{H}_i^{j^k} : x_0 \to c(x) * [F]_{ij}^k, \qquad (9)$$

де $\mathrm{H}_i^{j^k}$ — оператор парадигматизації, що діє в межах парадигматичного класу Π_k; індекси $i = 1, 2, \ldots, N; j = 1, 2, \ldots, p_j; k = 1, 2, \ldots, n_j$ використовуються відповідно для парадигматичних типів, граматичних класів та парадигматичних класів; N — кількість парадигматичних типів; P_j — кількість граматичних класів, які мають парадигматичний тип T_i; n_j — кількість парадигматичних класів у граматичному класі P_j; $[F]_{ij}^k$ — множина наборів квазіфлексій слів парадигматичного класу Π_k.

Оператор парадигматизації, який діє на множині слів кожного граматичного класу, визначається через оператори парадигматизації $\mathrm{H}_i^{j^k}$, що функціонують у межах своїх конкретних парадигматичних класів Π_k:

$$\mathrm{H}_i^j = \sum_{k=1}^{n_j} \mathrm{H}_i^{j^k} \bullet \delta(x; \Pi_k), \text{ де функція } \delta(x; \Pi_i) = \begin{cases} 1, x \in W(\Pi_k) \\ 0, x \notin W(\Pi_k) \end{cases}, \qquad (10)$$

де k, j, i — індекси парадигматичного, граматичного класів і парадигматичного типу.

Оператор парадигматизації H визначає відповідність її повної парадигми $[x]$ лексемі x. Алгоритмічна реалізація оператора H^{-1} лематизує певну словоформу, тобто забезпечує побудову вихідної форми слова.

Викладена формальна модель становить концептуальну основу для комп'ютерного моделювання та реалізації парадигматичних відношень у мові з флективним словозмінним компонентом.

Мова з флективним та аналітичним компонентами. Поняття парадигматичного класу, яке формально визначене формулами (4)–(10), узагальнимо щодо флективної мови з елементами аналітизму (тобто такої, в якій певним граматичним значенням відповідають аналітичні форми). Аналітична форма (АФ) може складатися з двох або більше компонентів, тобто $x = x_1 + x_2 + \ldots + x_Q$. При цьому кожний з компонентів АФ x_q, де $1 \le q \le Q$, може бути змінюваним або незмінюваним. У такому випадку аналітична форма має вигляд:

$$x = \sum_{q=1}^{Q} (c(x_q) * f(x_q)), \tag{11}$$

де $q = 1, 2, \ldots, Q$ — індекс компонента АФ, Q — кількість компонентів АФ. Якщо $Q = 1$, маємо один компонент, тобто x репрезентується за формулою (4). Складові елементи компонентів АФ такі: $c(x_q)$ — квазіоснова і $f(x_q)$ — квазіфлексія компонента АФ x_q. Квазіоснова $c(x_q)$ може набувати таких значень: $c(x_q) = 0$ (пустий рядок) у випадку суплетивних словозмінних форм; $c(x_q) = x_q$, якщо x_q — незмінюване слово; $c(x_q) = x_q - f(x_q)$. Квазіфлексія $f(x_q) = 0$, якщо компонент x_q є незмінюваним; $f(x_q) = x_q$, якщо всі словозмінні форми — суплетивні; в інших випадках квазіфлексією є деяка послідовність літер, що збігається з кінцевою частиною слова x_q.

Словозмінна парадигма відповідно до репрезентації (11) така:

$$\pi(x) = \{\sum_{q=1}^{Q} c(x_q) * \{f_j(x_q)\}\}, \tag{12}$$

де $j = 0, 1, 2, \ldots, n(T_i)$ — квазіфлексії компонента x_q у відповідних граматичних станах. При цьому будь-який граматичний стан може виражатися більш ніж однією формою (аналітичною або синтетичною).

Відношення парадигматизації π_i за наявності аналітичних форм визначаються у такий спосіб: два слова, що мають у словозмінній парадигмі аналітичні форми, належать до одного парадигматичного класу, якщо аналітичні форми будуються за однаковими схемами (мають ті самі допоміжні складники, а розрізняються лише смисловим компонентом), а відповідні компоненти АФ мають однакові набори квазіфлексій:

$$\begin{gathered} \forall\, x^1 = \sum_{q=1}^{Q} x_q^1,\ x^2 = \sum_{q=1}^{Q} x_q^2 \in P_i : x^1 \pi_i x^2 : \\ x^1 = \sum_{q=1}^{Q} c(x_q^1) * f_q^k, x^2 = \sum_{q=1}^{Q} c(x_q^2) * f_q^k,\ f_q^k \in [F]_q^k, \end{gathered} \tag{13}$$

де $[F]_q^k$ — набори квазіфлексій слів-компонентів АФ у відповідних граматичних значеннях. За відношенням парадигматизації отримуємо фактор-множину P_i / π_i, яка є множиною парадигматичних класів $\{\Pi_j\}$ граматичного класу P_i.

Таким чином, якщо словозмінна парадигма складається як із синтетичних, так і з аналітичних форм, до визначення парадигматичного класу додається ознака однаковості схем побудови аналітичних форм: АФ для відповідних граматичних значень будуються з однакових допоміжних складників (які можуть бути змінюваними); усі словозмінні складники АФ мають однакові набори квазіфлексій для відповідних граматичних форм.

Мова з аглютинативним компонентом. При моделюванні словозміни аглютинативних мов, зокрема турецької, ми виходимо з наявності у слова кореневого та афіксального складників:

$$W = [R] * [A], \tag{14}$$

де $[R]$ — «коренева» частина лексеми W (тобто та, яка містить корінь, щодо якого здійснюються словозмінні операції), а $[A]$ — його афіксальна частина; зірочкою позначено операцію конкатенації. У мові з аглютинативним словозмінним компонентом найінтенсивніші словозмінні процеси відбуваються насамперед в афіксальній частині за рахунок нарощування в певній послідовності афіксів, згідно із законом сингармонізму, відповідно до процесів спрощення і морфосемантизації. Це, звичайно, не виключає й інших словозмінних процесів, зокрема зумовлених варіаціями в кореневій частині $[R]$ при словозміні, особливо на межі між $[R]$ і $[A]$. Проте саме за рахунок афіксального нарощування та комбінування при словозміні реалізується процес породження системи граматичних значень — як щодо конкретної

лексеми, так і щодо лексичної системи в цілому. У своїй логічній структурі афіксальна словозміна веде до породження класифікаційних словозмінних схем.

Усю множину турецьких слів W, які належать до певного лексико-граматичного класу (наприклад, іменників), можна розбити на підмножини слів, які є еквівалентними. Ця еквівалентність (у подальшому — А-еквівалентність) визначається так: слова x та y, що належать до W, називатимемо А-еквівалентними, якщо вони при словозміні отримують тотожні набори афіксів та афіксальних комплексів (послідовностей).

Факт А-еквівалентності слів x та y позначатимемо символом $x\mathrm{A}y$. Формально факт А-еквівалентності можна подати таким чином.

Візьмемо два слова x та y з множини W і побудуємо їхні репрезентації у формі (14): $x = [R_x] * [A_x]$, $y = [R_y] * [A_y]$, де символами $[R_x]$ і $[A_x]$ позначено відповідно кореневу та афіксальну частини лексеми x, а $[R_y]$ і $[A_y]$ — відповідно кореневу та афіксальну частини лексеми y. Тоді з факту А-еквівалентності x та y, тобто з $x\mathrm{A}y$, випливає тотожна рівність афіксальних частин x та y: $x\mathrm{A}y \Rightarrow [A_x] \equiv [A_y]$.

Визначена у такий спосіб еквівалентність має такі формальні властивості:

1) рефлексивність: $x\mathrm{A}x$ (будь-яке слово є А-еквівалентним самому собі); 2) симетричність: $x\mathrm{A}y \Rightarrow y\mathrm{A}x$ (якщо слово x є А-еквівалентним слову y, то, очевидно, що у є А-еквівалентним слову x); 3) транзитивність: $x\mathrm{A}y$ та $y\mathrm{A}z \Rightarrow x\mathrm{A}z$ (якщо слово x — А-еквівалентне слову y, а y — А-еквівалентне z, то x — А-еквівалентне z).

Позначимо символом $K^{\mathrm{A}}(x)$ клас слів з W, які є А-еквівалентними слову x. Якщо взяти два класи — $K^{\mathrm{A}}(x)$ та $K^{\mathrm{A}}(y)$, тобто класи слів, А-еквівалентних словам x та y відповідно, то можливими є тільки два взаємовиключні варіанти:

$$1)\ K^{\mathrm{A}}(x) \equiv K^{\mathrm{A}}(y) \text{ або } 2)\ K^{\mathrm{A}}(x) \cap K^{\mathrm{A}}(y) = \varnothing. \qquad (15)$$

Це означає, що ці два класи або повністю збігаються, або не мають жодного спільного елемента (їх перетин є порожньою множиною).

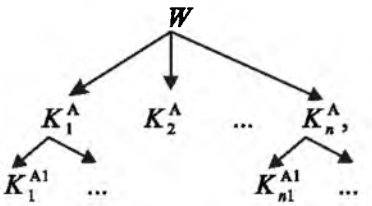
Отже, вся множина слів W подається у вигляді об'єднання класів А-еквівалентності, взаємний перетин яких є порожнім:

$$W = \cup K^{\mathrm{A}}(x);\ K^{\mathrm{A}}(x) \cap K^{\mathrm{A}}(x') = \varnothing \text{ при } x \neq x',\ x, x' \in W^{\mathrm{A}}, \qquad (16)$$

де W^{A} — певна підмножина А-нееквівалентних слів з W.

Слід зазначити, що множини, виражені формулами (15)–(16), насправді задають формальне визначення поняття класифікації безвідносно до того, які конкретно об'єкти розглядаються. Якщо є множина W об'єктів x ($x \in W$) будь-якої природи і між ними встановлено яке-небудь рефлексивне, симетричне і транзитивне відношення А, то воно породжує певну класифікацію елементів множини W. Характерною її ознакою є інваріантний поділ початкової множини об'єктів на підмножини, які не перетинаються, — класи еквівалентності відповідно до принципу А, що зумовлює такий поділ. До кожного класу еквівалентності потрапляють елементи з подібними властивостями, у той час як властивості елементів з різних класів дещо різняться. У свою чергу, в кожному з класів $K^{\mathrm{A}}(x)$ (або в певній їх сукупності) може бути встановлена своя, «дрібніша» класифікація шляхом визначення нових відношень еквівалентності, і цей процес індукування більш тонких класифікацій на підмножинах, у принципі, може продовжуватися.

Графічно класифікаційну схему можна представити у вигляді ієрархічного графа:



(17)

де W — певна вихідна множина об'єктів; K_1^A, K_2^A, ..., K_n^A становлять перший рівень класифікаційної ієрархії за принципом класифікації А; K_1^{A1}, ..., K_{n1}^{A1} — другий рівень класифікаційної ієрархії за принципом класифікації А1 і т. ін.

Технологія побудови граматичних лексикографічних систем, їхня структура та функції. Розглянемо технологію створення ГЛС на прикладі російської мови. Формування граматичної лексикографічної бази даних (ЛБД) російської мови здійснювалося на основі «Грамматического словаря русского языка» А. А. Залізняка (ГСЗ) [5], який достатньо повно моделює словозмінну систему російської мови. На першому етапі було переведено книжковий варіант словника в електронну форму засобами сканування та розпізнавання тексту. Цифровий варіант тексту ГСЗ (880 сторінок) збережено в doc-форматі, після чого було відкоректовано електронний текст. На наступному етапі здійснено конверсію з doc-формату в html-формат із системою кодування Unicode засобами текстового редактора Microsoft Word. Для автоматичної конверсії електронного тексту ГСЗ в ЛБД розроблено програмне забезпечення виділення елементів структури ГСЗ відповідно до будови лексикографічної системи і з використанням поліграфічних ознак їх текстової ідентифікації.

Структуру даних репрезентовано реляційною моделлю. На рис. 3 подано схему зв'язків між таблицями граматичної ЛБД російської мови: таблицею реєстрових слів (nom), граматичних класів (Parts), парадигматичних класів (indent), квазіфлексій (flex), словозмінних типів (gr) та акцентуаційних класів (accent).

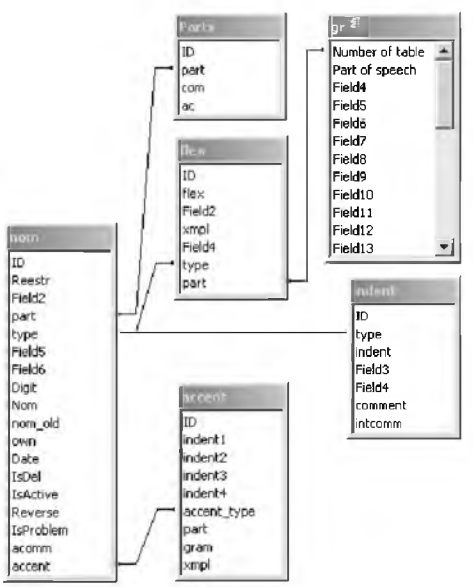


Рис. 3. Схема зв'язків між таблицями граматичної ЛБД російської мови.

[5] *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение.—3-е изд.— М., 1987.— 880 с.; 4-е изд., испр. и доп.— М., 2003.— 800 с.

Для супроводу граматичних ЛБД відповідних мов розроблено інструментальні комплекси, які працюють під управлінням операційної системи Microsoft Windows XP/2003/2008 або Windows Vista. Створені граматичні ЛБД функціонують під СУБД Microsoft SQL Server 2008. Комплекси орієнтовано на роботу в мережевому середовищі.

Розглянемо функціональні можливості інструментального комплексу, який розроблено для супроводу ЛБД російської мови. У розробленні програмного інтерфейсу було використано елементи управління операційного середовища Windows. Головне вікно програми розділено на три зони: функціональну, реєстрову та зону лексикографічної інформації.

Функціональна зона складається з таких підзон: загальне меню, інструментарій для редагування, інструментарій для виконання запитів на мові SQL, інтерфейс для пошуку слів.

Загальне меню містить пункти «Файл», «Вид», «Словник», «Загальний вибір», «Вибірка» і «Довідка». Кожний з цих пунктів складається з підпунктів, завдяки яким реалізуються функції, що надаються користувачеві, а саме: перегляд реєстру; отримання повної словозмінної парадигми обраного з реєстру слова і його основних граматичних характеристик (ця інформація подається в зоні лексикографічної інформації); виведення та перегляд частини реєстру (за певною ознакою або за комбінацією ознак, наприклад за лексико-граматичним класом, номером парадигматичного класу тощо); вибірка лексичних омонімів, власних назв, тощо; виведення кількісних характеристик щодо наповнення обраної групи слів (парадигматичних класів, частин мови, омонімів тощо); пошук слів у реєстрі; побудова прямого або інверсійного словника (встановлення прямого або інверсійного сортування в реєстрі); додавання нових слів, редагування реєстрових слів, видалення слів з реєстру.

Наступними важливими функціями системи є виконання операцій з парадигматичними класами: додавання, редагування, видалення парадигматичних класів, додавання та редагування квазіфлексій синтетичних форм і типів процедур утворення аналітичних форм (для аналітичних або аналітико-синтетичних мов); побудова словника квазіоснов (для мов флективного типу; словник квазіоснов використовується програмами морфологічного та синтаксичного аналізу).

Дефектність та варіативність словозмінної парадигми. Дослідження типів дефектності та варіативності словозмінних парадигм можливе завдяки розробленим в УМІФ НАН України потужним лексикографічним ресурсам — граматичним лексикографічним системам. Продемонструємо реалізацію можливості такого дослідження на прикладі російської мови на основі її електронного граматичного словника (ЕГС), який призначений насамперед для автоматичного опрацювання текстів. Цей ресурс можна також використовувати й для дослідження властивостей словозмінних парадигм, таких як її дефектність (неповнота) і варіативність.

В основу формування лексикографічної бази ЕГС покладено визначення парадигматичного класу (ПК), який об'єднує лексеми з однаковими правилами утворення їхньої словозмінної парадигми щодо кількості словозмінних форм у парадигмі, набору квазіфлексій, а також схеми наголосу.

Сформована ЛБД містить 170 тис. словникових одиниць, з них 100 тис. — реєстр словника А. А. Залізняка, 70 тис. — уведені нами в автоматизованому режимі нові слова.

Дослідження виконувалося на матеріалі 818 парадигматичних класів іменників, 169 класів ад'єктивів та 966 дієслівних класів.

Дефектну парадигму слова розуміємо як словозмінну парадигму з відсутніми в ній словоформами для деяких граматичних значень. Як приклад можна навести дефектну словозмінну парадигму російського іменника *лёт* з відсутнім у нього значенням множини: *лёт, лёта, (с) лёту, лёту, лёт, лётом, лёте, (на) лету́*.

Для опису типів дефектності словозмінних парадигм нами введено параметр дефектності, який є переліком словозмінних форм, не реалізованих у парадигмі. У наведеному прикладі параметр дефектності представлено множиною, яка складається з 6-ти елементів: *def* = {7, 8, 9, 10, 11, 12} [6]. Параметри дефектності, що розрізняються, відповідають різним типам дефектності.

Словозмінну парадигму, в якій хоча б одна граматична форма реалізується більш ніж однією словоформою, називаємо варіативною. Опис варіативних типів парадигм здійснюється за допомогою параметра варіативності, що його представляє множина пар чисел, у кожній з яких перше число — це номер граматичного значення в парадигмі, а друге — кількість словоформ, що відповідають конкретному граматичному значенню. У нашому прикладі параметр варіативності *var* = {<2,2>, <6,2>} означає, що лексема *лёт* у родовому та місцевому відміках однини має дві форми.

Типи дефектності та типи варіативності визначаються шляхом аналізу таблиць квазіфлексій, які в структурі бази даних описують парадигматичні класи конкретних частин мови. Аналіз виконано в автоматичному режимі за допомогою спеціально розроблених програмних модулів. У таблиці квазіфлексій описано 2015 ПК.

Виявлено 1207 ПК з дефектністю певного типу, 808 парадигматичних класів мають словозмінну парадигму, що включає всі теоретично можливі форми, тобто в цих класах слів дефектність відсутня: *def* = 0. (Дані про кількість дефектних парадигматичних класів у межах кожної частини мови див. у табл. 1).

Таблиця 1

Кількість дефектних парадигматичних класів

Частини мови	Кількість ПК	Кількість ПК з дефектністю
іменники	818	188
ад'єктиви	138	46
дієслова	966	966
займенники-іменники	37	5
займенники-прикметники	31	2
кількісні числівники	25	0
Разом	2015	1207

Найбільшу кількість типів дефектності парадигми виявлено в дієслів (69), а в кількісних числівників вона відсутня. Повна дієслівна парадигма за прийнятим у статті визначенням складається з 49 словозмінних форм. До них, крім особових форм дієслова, належать форми дієприкметників та дієприслівників, що по-різному реалізуються в дієслів доконаного та недоконаного виду. Крім того, майже в усіх дієсловах недоконаного виду відсутні синтетичні форми майбутнього часу. Саме комбінації номерів, які відповідають цим дієслівним формам, становлять більшу частину типів дефектності дієслова.

Дефектність ад'єктивних парадигм стосується в більшості випадків можливості (неможливості) прикметників утворювати короткі форми (всі або деякі).

[6] Такий запис означає, що у словозмінній парадигмі не реалізуються множинні форми в усіх відмінках.

Серед типів дефектності парадигм іменників найширше представлені два типи, які описують ситуації pluralia tantum (1110 слів) та singularia tantum (8865 слів). Інші 13 типів стосуються одиничних випадків неможливості утворювати певні відмінкові форми. Нижче наводимо результати аналізу типів дефектності словозмінних парадигм іменників та ад'єктивів (див. табл. 2 і табл. 3).

Таблиця 2

Типи дефектності парадигм іменників

Код ТД	Тип дефектності (ТД)	Кількість ПК з даним ТД	Кількість слів з даним ТД	Слово
d_1	{1, 2, 3, 4, 5, 6}	66	1110	грабли
d_2	{1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12}	1	3	щец
d_3	{1, 2, 3, 4, 5, 6, 8}	1	1	бразды
d_4	{1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12}	1	2	тары-бары
d_5	{1, 2, 3, 4, 5, 6, 8, 9, 11, 12}	1	1	полсуток
d_6	{2, 3, 5, 6, 7, 8, 9, 10, 11, 12}	7	34	пламень
d_7	{2, 3, 6, 7, 8, 9, 10, 11, 12}	1	1	полымя
d_8	{2, 5, 6, 7, 8, 9, 10, 11, 12}	1	1	теля
d_9	{3, 4, 5, 6, 7, 8, 9, 10, 11, 12}	1	1	ведомо
d_{10}	{3, 4, 5, 7, 8, 9, 10, 11, 12}	1	1	полдороги
d_{11}	{3, 5, 7, 8, 9, 10, 11, 12}	2	2	полслова
d_{12}	{7, 8, 9, 10, 11, 12}	92	8865	богочеловек
d_{13}	{7, 9, 10, 11, 12}	1	1	зло
d_{14}	{8}	7	40	башка
d_{15}	{8, 10}	5	23	брюзга

Таблиця 3

Типи дефектності парадигм прикметників

Код ТД	Тип дефектності	Кількість ПК з даним ТД	Кількість слів з даним ТД	Слово
d_1	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24}	3	47	рад, должен
d_2	{25}	2	24	полубольной
d_3	{25, 26}	1	1	взрослый
d_4	{25, 26, 27}	1	187	поблёклый
d_5	{25, 26, 27, 28}	26	15143	меньшой
d_6	{26}	5	22	сверхдолгий
d_7	{26, 27}	1	23	навислый
d_8	{26, 27, 28}	1	2	полутяжёлый

Повна парадигма ад'єктивів включає 28 словозмінних форм. Лінгвістична інтерпретація виявлених типів дефектності ад'єктивної парадигми така: d_1 — словозмінна парадигма складається тільки з коротких форм, d_2 — у парадигмі відсутня коротка форма чоловічого роду, d_3 — відсутні короткі форми чоловічого та жіночого роду, d_4 — немає коротких форм чоловічого, жіночого та середнього роду, d_5 — у парадигмі відсутні короткі форми, d_6 — відсутня коротка форма жіночого роду, d_7 — немає коротких форм жіночого та середнього роду, d_8 — у словозмінній парадигмі відсутні короткі форми жіночого та середнього роду, а також коротка форма множини.

Результати аналізу типів варіативності (ТВ) словозмінних парадигм вносилися в таблицю, яка містить код ТВ, код ТД, тип варіативності, що описується параметром var, кількість парадигматичних форм (ПФ), кількість ПК, які мають

даний тип варіативності, приклад слова з даним ТВ. Нижче наводимо фрагмент таблиці ТВ прикметників (див. табл. 4).

Таблиця 4

Типи варіативності прикметників

Код ТВ	Код ТД	Тип варіативності	Кіл-ть ПФ	Кіл-ть ПК	Слово
v_2	d_0	{<4,2>, <22,2>}	30	40	больной
v_3	d_0	{<4,2>, <22,2>, <28,2>}	31	20	строгий
v_{16}	d_5	{<2,2>, <3,2>, <4,3>, <14,2>, <15,2>, <22,2>}	31	1	дочерний
…	…	…	…	…	…
v_{20}	d_8	{<4,2>, <22,2>}	27	1	полутяжёлый

Отримані результати дослідження варіативності дозволяють твердити, що в російських дієслів варіативність виявляється найчастіше в дієприслівникових формах (*вы́меривши // вы́мерив*), у формах наказового способу (*вы́меряй // вы́мери // вы́мерь*), в особових формах майбутнього часу дієслів доконаного виду (*вы́меряю // вы́мерю* і т. д.), в особових формах теперішнього часу дієслів недоконаного виду (*щеплю́ // щепа́ю* і т. д.), у дієприкметникових формах (*ще́плющий // щепа́ющий, ще́племый // щепа́емый*). У прикметників варіативність виявляється частіше у формах знахідного відмінка однини чоловічого роду (*стро́гий // стро́гого*), знахідного відмінка множини *(стро́гие // стро́гих*), у коротких формах (*стро́ги // строги́, шустёр, шустр*). Рідше зустрічається варіативність в інших формах непрямих відмінків, наприклад родового та давального відмінка однини середнього роду (*куку́шкиного // куку́шкина*; *куку́шкиному // куку́шкину*). У іменників варіативність виявляється в усіх відмінкових формах. Усього для іменників визначено 59 різних ТВ. Найбільшою варіативністю відзначається іменник чоловічого роду *мох*, у парадигмі якого 26 форм, причому варіативність наявна в усіх відмінках, крім називного та знахідного в однині. А найбільшу кількість варіантів словозмінних форм мають родовий (*мха, мху, мо́ха, мо́ху*) та місцевий (*мхе, мо́хе, мху, мо́ху*) відмінки однини.

Дослідження типів дефектності та варіативності словозмінної парадигми дозволило здійснити додаткову параметризацію словникових одиниць, приписавши їм значення ТД і ТВ. Отримані дані можуть бути застосовані в типологічних дослідженнях словозмінних систем флективних мов.

T. P. LYUBCHENKO, I. V. SHEVCHENKO, K. V. SHYROKOV

GRAMMATICAL LEXICOGRAPHICAL SYSTEMS WITH THE FLECTIONAL AND AGGLUTINATIVE COMPONENT OF INFLECTION

Grammatical lexicographical systems (GLS) are being developed in the Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine for different languages within the framework of research works to create the National dictionary database of Ukraine. The article presents a conceptual model of inflectional systems of languages with the flectional component in inflection, those with the flectional and the analytical component, and those with the agglutinative component. The technology of construction of GLS, their structure and functions are described. In addition to its main purpose (use in the automatic text processing) the GLS are also used in the study of inflectional systems of respective languages. The article, in particular, deals with the results of research on the parameters of incompleteness and variability of the inflectional paradigm in Russian. The data obtained can be used in the typological studies of the inflectional systems of the inflected languages.

Key words: grammatical lexicographical system, grammar dictionary, paradigmatic type, incompleteness of inflection paradigm, variability of inflection paradigm.