УДК 658.012.011.56

Т.П. Любченко

ПРОГРАМНО-ТЕХНОЛОГІЧНІ АСПЕКТИ СТВОРЕННЯ ГРАМАТИЧНИХ ЛЕКСИКОГРАФІЧНИХ СИСТЕМ

Розглянуто комп'ютерну технологію укладання електронних граматичних словників (ЕГС). Описано структуру лексикографічної бази даних (ЛБД) ЕГС та принципи її побудови. Запропоновано клієнтську програму для редагування ЛБД ЕГС, розглянуто її функціональні можливості.

Вступ

Електронні граматичні словники розробляються як частина інтегрованої лексикографічної системи Українського мовно-інформаційного фонду НАН України (УМІФ НАН України). В основу розробки покладено теорію лексикографічних систем [1–4].

ЕГС розробляються у першу чергу для мов, які включено до системи багатомовного машинного перекладу (МП): української, російської, англійської, німецької, іспанської, французської та турецької мов. Словники орієнтовано на писемні варіанти мов. ЕГС призначені, насамперед, для використання їх у системах автоматичного опрацювання текстів (наприклад, для аналізу, лематизації та синтезу словоформ, на етапах морфологічної розмітки тексту, тощо). Крім цього, передбачено надання користувачеві можливості доступу до словника як до довідково-інформаційної системи (пошук слів, надання інформації щодо словозміни певних реєстрових одиниць).

Слід зазначити, що в останні десятиліття багатьма дослідниками розроблено системи морфологічного аналізу, синтезу, лематизації та виправлення помилок як для російської мови (наприклад, [5–9]), так і для багатьох інших мов світу. Зрозуміло, що більшість систем такого типу є мовнозалежними. Отже відсутність універсальної морфологічної моделі (граматичної Л-системи) спонукала нас до спроби розробити загальні (єдині) принципи побудови граматичних Л-систем (для певного класу флективних мов).

Принципи побудови словозмінної моделі лексики

Для побудови граматичного словника флективної мови визначальним фактором є наявність формальної моделі словозміни, що означає встановлення та формалізацію лінгвістичних критеріїв, згідно з якими вся множина слів мови розбивається на певні підмножини, взаємний перетин яких є порожнім, а всередині кожної з них словозміна відбувається за однаковими правилами. Підмножини слів з такими властивостями називаються словозмінними парадигматичними класами.

Моделювання розподілу множини слів мови на словозмінні парадигматичні класи відбувається у декілька етапів. На першому визначається поняття парадигматичного типу, в означенні якого принципову роль відіграють поняття граматичної категорії, граматичного значення та граматичної форми [10].

Уведемо позначення. Нехай L – фіксована мова, W – сукупність слів мови L ;

P_j , (j = 1, 2, ..., p) – граматичні класи[1], p – кількість граматичних класів для даної мови;

$W(P_j)$ – множина слів мови L, яка належить граматичному класу P_j;

T_i , (i = 1, 2, …, N) – парадигматичні (морфологічні) типи, N – кількість парадигматичних (морфологічних) типів;

$W(T_i)$ – множина слів мови L, яка належить типу T_i,

$\Omega(T_i)$ – множина граматичних значень, що відповідають типу T_i.

[1] граматичний клас – аналог частини мови

За ознакою приналежності до певної частини мови та за додатковими ознаками, які є класифікуючими (не словозмінними) у межах певної частини мови, множина слів W певної мови L розподіляється на підмножини, котрі будемо називати граматичними класами.

Отже,

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

На цьому етапі викладу вважатимемо омонімію знятою, а омоніми промаркованими. Тоді $W(P_k) \cap W(P_l) = \varnothing$ при $k \neq l,\ k,l = 1,2,...,p$ (взаємний перетин слів із різних граматичних класів є порожньою множиною).

За словозмінними категоріями, що визначають словозмінну парадигму конкретних слів (сукупність граматичних значень та відповідних граматичних форм), вводяться парадигматичні типи (ПТ). Так, що $W = \bigcup_{i=1}^{N} W(T_i)$.

Докладний опис парадигматичних типів кожної з розглядуваних мов можна знайти в наших публікаціях [11, с. 128–156, 218–223], [12]. Тут для прикладу наведемо лише парадигматичні типи для російської мови (табл. 1).

Принцип розподілу лексики флективної мови на парадигматичні типи, граматичні класи та парадигматичні класи показано на рис. 1.

Парадигматичні типи можуть бути притаманними декільком граматичним класам, усередині кожного з яких виділяються парадигматичні класи. Як видно зі схеми, згідно з принципом розподілу

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

де $P_j \subseteq T_i$, p_i – кількість граматичних класів, в яких словозміна має парадигматичний тип T_i;

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

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

Отже, $W(T_i) = \bigcup_{j=1}^{p_i} (\bigcup_{k=1}^{n_j} W(\Pi_k))$.

Таблиця 1. Парадигматичні типи російської мови

Парадигматичний тип	Граматичні класи	Граматичні категорії, що визначають словозміну	Кількість граматичних значень у повній парадигмі
Субстантивний (T^S)	Іменники, займенники-іменники	число, відмінок	12
Ад'єктивний (T^A)	Прикметники, займенники-прикметники, порядкові числівникі, дієприкметники	рід, число, відмінок	28
Дієслівний парадигматичний тип (T^V)	Дієслова доконаного виду, дієслова недоконаного виду, двовидові дієслова	стан, час, число, особа, спосіб, рід	46
Парадигматичний тип кількісних числівників (T^C)	Кількісні числівники	відмінок	6
«Нульовий» парадигматичний тип (T^0) – незмінювані слова	Прислівники, вигуки, сполучники, частки, прийменники, предикати	–	1

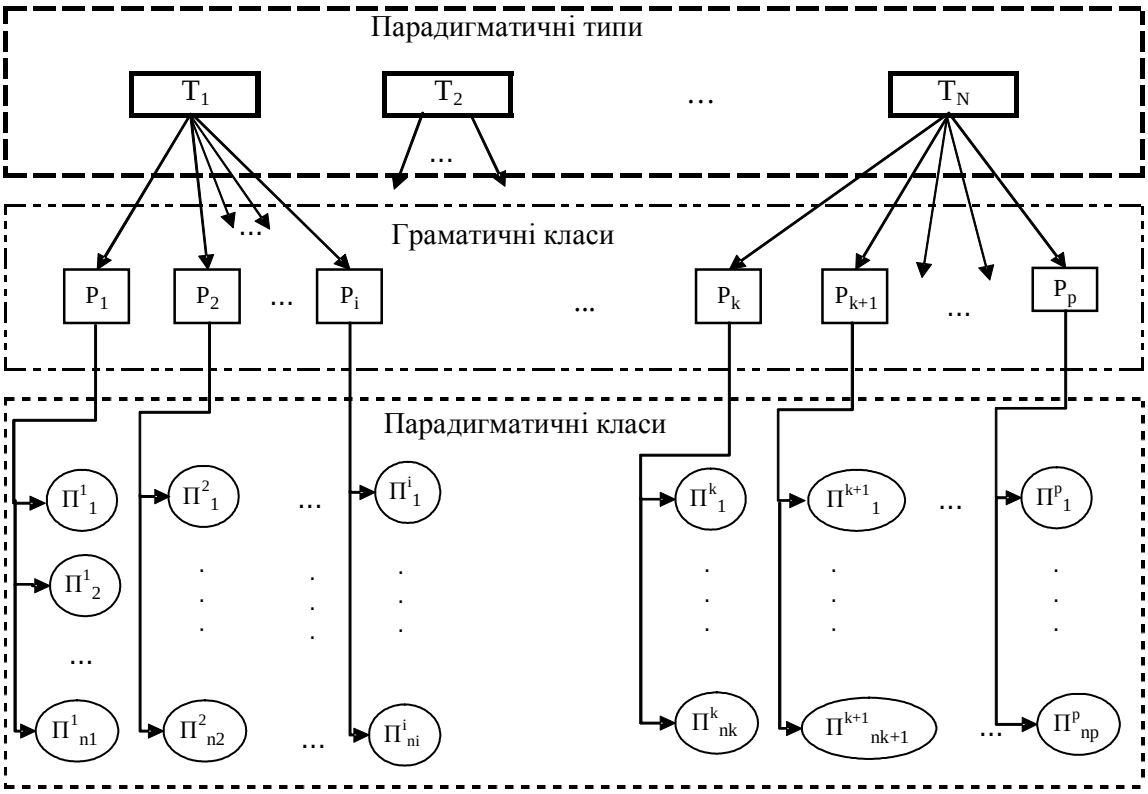

Рис. 1. Схема розподілу лексики флективної мови

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

$$x = c(x) * f(x), \qquad (4)$$

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, в російській мові 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, нулю дорівнює квазіфлексія.

Повна словозмінна парадигма $[x]$ слова x, що належить до граматичного класу P_j (парадигматичного типу T_i) , представляється у вигляді

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

де $f_j(x)$, $j = 0,1,2,..., n(T_i)$ – 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 $\nu(w_j(x))$, який задається цілим числом, рівним кількості можливих форм лексеми x у граматичному значенні ω_j. У загальному випадку:

$$f_j(x) = \bigcup_{l=0}^{\nu(w_j(x))} f_{jl}, \qquad (6)$$

$l = l(j) = 0,1,2, ...$ – індекс кількості словоформ у граматичному значенні ω_j;

$f_0(x)$ – квазіфлексія початкової форми, яка для іменника 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 значень для парадигматичного типу T_i.

Покладемо

$$F = \bigcup_{x \in W} (\{ f_0(x) \}, \{ f_{1l}(x) \}, \ldots ,$$

$$\ldots \{ f_{n(T_i)l}(x) \}) \equiv \{ f_{jl}{}^1, f_{jl}{}^2, \ldots, f_{jl}{}^{N_i} \}, \quad (7)$$

$$j = 0,1,2,\ldots, n(T_i), \quad l = l(w_j) = 0,1,2,\ldots$$

Тоді

$$F = \bigcup_{k=1}^{N_i} [F]^k , \quad (8)$$

де

$$[F]^k = \{f^k\} = \{f_{jl}{}^k, j = 0,1,\ldots n(T_i)\}. \; N_i = N(T_i), l = l(w_i).$$

Таким чином, кожна множина $[F]^k$ складається з квазіфлексій слів, які мають у всіх своїх граматичних значеннях ω_1, ω_2, …, $\omega_{n(Ti)}$ (парадигматичного типу T_i) однакові змінні складові.

Оскільки $[F]^k$ побудовані таким чином, що в них увійшли унікальні набори квазіфлексій, тобто $[F]^i \neq [F]^j$ при $i \neq j$ $(i, j, = 1,2,\ldots, N_i)$, то для кожного граматичного класу P_i (парадигматичного типу T_i) можна побудувати відношення π_i на декартовому добутку $P_i \times P_i$, яке визначається так:

$$\forall x^1, x^2 \in P_i \;\; x^1 \pi_i x^2 : \; x^1 = c(x^1) * f^k,$$
$$x^2 = c(x^2) * f^k, \; f^k \in [F]^k. \quad (9)$$

Це відношення є відношенням еквівалентності, оскільки воно, очевидно, є рефлексивним, симетричним та транзитивним. Назвемо його відношенням парадигматизації.

Фактор-множина P_i / π_i – множина парадигматичних класів граматичного класу P_i (парадигматичного типу T_i). Очевидно, що різні словозмінні парадигматичні класи не перетинаються. Отже P_i – об'єднання парадигматичних класів: $P_i = \bigcup_{j=1}^{n} \Pi_j$. До одного парадигматичного класу входять тільки ті слова, які мають однакові набори квазіфлексій для всіх граматичних форм, а відрізняються один від одного лише незмінною складовою $c(x)$. Зрозуміло також, що слова з одного класу еквівалентності, визначеного в такий спосіб, мають і однакові правила словозміни.

Таким чином, для кожного з граматичних класів (парадигматичних типів T_i) будується розбиття на множини слів, що не перетинаються, і які є парадигматичними класами, всередині кожного з яких діють єдині правила словозміни. Для мов флективного типу це означає однаковість флексій граматичних форм та збіг характеру чергування в основі.[2]

Для автоматичної побудови повної парадигми за вихідною (початковою) формою x_0 визначається оператор парадигматизації

$$\mathrm{H} : x_0 \to [x] = c(x) * \{ f_0(x), f_1(x), \ldots$$
$$\ldots, f_n(x) \} \equiv \{ c(x) * f_0(x), c(x) * \quad (10)$$
$$* f_1(x), \ldots, c(x) * f_n(x) \},$$

дія якого визначається відношенням $\pi(x^1, x^2)$.

Оператор повної парадигматизації (який діє на множині лексем W) визначається за формулою

$$\mathrm{H} = \sum_{i=1}^{N} \mathrm{H}_i \cdot \delta(x; T_i) , \quad (11)$$

де

$$\delta(x; T_i) = \begin{cases} 1, & x \in W(T_i), \\ 0, & x \notin W(T_i), \end{cases} \quad (12)$$

N – кількість парадигматичних типів, H_i – оператор парадигматизації, який діє на множині лексем відповідного парадигматичного типу $W(\mathrm{T}_i)$ [3]:

$$\mathrm{H}_i : x_0 \to [x] \; \forall x \in W(T_i) . \quad (13)$$

[2] флексія слова плюс частика основи із чергуванням складають квазіфлексію; частина основи, яка є однаковою в усіх словоформах складають квазіоснову

[3] на множині лексем кожного з парадигматичних типів діє свій оператор парадигматизації. Перш за все тому, що кожен із парадигматичних типів характеризується своїм комплексом значень граматичних категорій, котрий відрізняється для кожного з ПТ.

Оскільки має місце (2), можна стверджувати (записати), що

$$\mathrm{H}_i = \sum_{j=1}^{p} \mathrm{H}_i^j \cdot \delta(x; P_j), \qquad (14)$$

де

$$\delta(x; P_j) = \begin{cases} 1, & x \in W(P_j), \\ 0, & x \notin W(P_j), \end{cases} \qquad (15)$$

$i = 1,2,\ldots,N$; p – кількість граматичних класів у множині $W(T_i)$.

Оператор H_i^j діє на множині лексем у межах граматичного класу P_j парадигматичного типу T_i:

$$\mathrm{H}_i^j : x_0 \to [x] \quad \forall x \in W(P_j) \subseteq W(T_i), \qquad (16)$$

У свою чергу, оскільки P_j є об'єднанням парадигматичних класів ($P_j = \bigcup_{k=1}^{n} \Pi_k$), можна записати, що

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

$$\mathrm{H}_i^j = \sum_{k=1}^{n_j} \mathrm{H}_i^{jk} \cdot \delta(x, \Pi_k), \qquad (17)$$

де H_i^{jk} – оператор парадигматизації, який діє у межах парадигматичного класу Π_k;
$i = 1,2,\ldots,N$; $j = 1,2,\ldots,n_j$; n_j – кількість парадигматичних класів у множині $W(P_j) \subseteq$ $\subseteq W(T_i)$,
$[F]_i^k$ – множина наборів квазіфлексій слів граматичного класу P_j i-го парадигматичного типу (T_i); функція

$$\delta(x; \Pi_k) = \begin{cases} 1, & x \in W(\Pi_k), \\ 0, & x \notin W(\Pi_k). \end{cases} \qquad (18)$$

Таким чином для кожного з парадигматичних типів $W(T_i)$ оператор парадигматизації визначається незалежно.

Оператор H відображає лексему x на її повну парадигму $[x]$. Його реалізація для лексики відповідної флективної мови відбувається за допомогою словника квазіфлексій і набору алгоритмів побудови повних словозмінних парадигм. За допомогою парадигматичного словника довільній лексемі приписується її словозмінний тип, далі за алгоритмом морфологічного аналізу (з використанням набору алгоритмів побудови повних словозмінних парадигм) здійснюється граматична ідентифікація лексеми x. Після цього лексема набуває представлення (4).

Алгоритмічна реалізація оператора H^{-1} здійснює процес лематизації, тобто зведення довільної словоформи до її вихідної канонічної форми.

Вищевикладена морфологічна модель складає концептуальну основу для комп'ютерного моделювання та реалізації функції парадигматичних відношень для певного класу флективних мов. Згідно з викладеною концепцією будова (структура) парадигматичної лексикографічної системи для флективної мови показана на рис. 2.

У наведеній схемі елементи мають таку інтерпретацію:

V(PAR) – множина словникових статей граматичної Л-системи;

$\Lambda(\mathrm{PAR}) = \mathrm{F}^{\mathrm{PAR}}\mathrm{V}(\mathrm{PAR}) = \{x_0\}$;

$\mathrm{P}(\mathrm{PAR}) = \mathrm{C}^{\mathrm{PAR}}\mathrm{V}(\mathrm{PAR}) = \{[x]\}$;

$\mathrm{H} = \mathrm{H}^{\mathrm{PAR}}$ – оператор парадигматизації: $\mathrm{H}^{\mathrm{PAR}}x_0 = [x]$;

$\mathrm{H}^{-1} = \mathrm{H}^{\mathrm{LEM}}$ – оператор лематизації:

$\mathrm{H}^{\mathrm{LEM}}\chi(x) = x_0$, де $\chi(x)$ – будь-який елемент парадигми $[x]$.

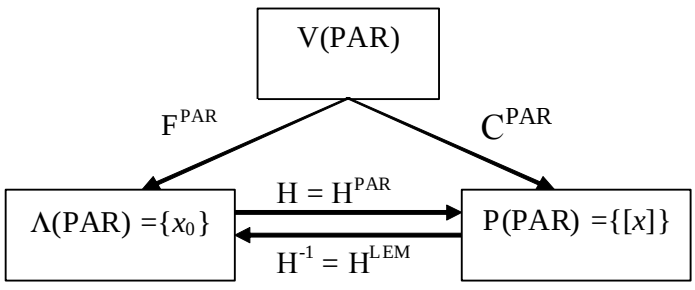

Рис. 2. Структура парадигматичної Л-системи

Технологія створення комп'ютерної лексикографічної бази даних електронного граматичного словника

Технологію створення ЛБД ЕГС розглянемо на прикладі російської мови. Джерелом первинної лінгвістичної інформації зі словозміни російської мови є „Грамматический словарь русского языка" А.А. Залізняка (у подальшому – ГСЗ) [13], який з достатньою повнотою моделює словозмінну систему російської мови.

Технологія створення ЛБД російського граматичного словника включає такі етапи:

- переведення тексту ГСЗ з паперової до електронної форми (сканування);
- коректура відсканованого тексту;
- розробка структури лексикографічної системи ГСЗ, мови її розмітки та встановлення ідентифікаторів елементів структури;
- автоматична конверсія електронного тексту ГСЗ в ЛБД відповідно до розробленої структури;
- розробка алгоритмів парадигматичної класифікації ГСЗ та їх програмна реалізація;
- формування парадигматичної ЛБД.

На основі цієї технології було створено ЕГС російської мови [14], який є складовою інтегрованої граматичної Л-системи може використовуватись як довідково-інформаційна система користувачами-філологами, які працюють з російськими текстами.

Джерела лінгвістичної інформації для створення граматичних словників німецької, англійської та іспанської мов використано граматики і словники відповідних мов [15–19].

Структура лексикографічної системи ЕГС

На внутрішньому рівні архітектури лексикографічної системи ЕГС (Л-системи ЕГС) структура лінгвістичних даних репрезентується реляційною моделлю, в якій визначено таку множину таблиць:

- таблиця **nom** реєстрових одиниць Reestr, де для кожної реєстрової одиниці зазначено код лексико-граматичного класу part та номер парадигматичного класу (поле type);
- таблиця квазіфлексій **flex**, в якій для кожного граматичного значення (поле NumbOfGrForm) кожного з парадигматичних класів (поле type) задані відповідні квазіфлексії flex;
- таблиця **indent**, що задає параметри та характеристики, які є однаковими для кожного з парадигматичних класів;
- таблиця **Parts** лексико-граматичних класів із відповідними їм кодами;
- таблиця **gr** словозмінних типів.

Зв'язок між таблицями **nom**, **flex**, **indent** здійснено за номером парадигматичного класу (поле type); а між таблицями **nom**, **Parts**, **gr** – за полем part, що зображено на рис. 3.

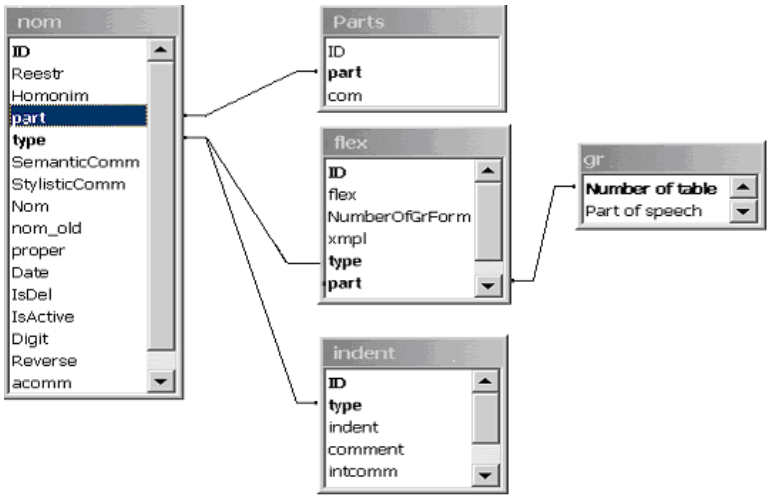

Рис. 3. Схема зв'язків між таблицями даних (російська або українська мова)

Наведено структуру Л-системи ЕГС розроблено для української та російської мов, в яких словозмінна парадигма представляється синтетичними формами.

При розробці моделі даних для інших мов (предметом нашого розгляду крім російської та української мови були англійська, іспанська та німецька) з метою уніфікації представлення їх в ЛБД, було використано підхід, аналогічний до вищенаведеного, тобто за основну одиницю словника обрано квазіоснову. Відмінність полягає лише в тому, що в ЛБД відзначених мов до опису парадигми крім квазіфлексій додаються ще й типи процедур, які використовуються при побудові тих чи інших аналітичних форм.

В англійській та іспанській мовах квазіоснова правильних дієслів, більшості іменників та прикметників збігається з вихідною формою слова. Представлення словникової одиниці у вигляді квазіоснови стосується тільки неправильних дієслів й незначної кількості іменників.

Німецька мова характеризується більш складними словозмінними процесами: прості (синтетичні) форми утворюються в ній флективним способом, а складні (аналітичні) – за певними схемами (тобто процедурно), причому словозмінними в аналітичних формах є як основний змістовий компонент, так і допоміжний; наявна велика кількість чергувань; має місце явище відокремлюваності префіксів у певної групи дієслів, тощо. Наявність великої кількості чергувань, що виникають у словозмінних процесах слів німецької мови, висунула вимогу, по-перше, урахування цього факту при розбитті множини словозмінних одиниць мови на парадигматичні класи, і, по-друге, введення до структури даних, що описують словозмінну парадигму, додаткових полів.

До таблиці **indent** додатково введено такі поля: поле PosAlter – містить вказівку на позицію символа в слові, з якого починається сегмент, у якому відбуваються зміни; поле QuantAlter – кількість змінюваних символів; поле outstr – послідовність символів, яка має замістити позначений сегмент; поле IsTrent – ознака про наявність/відсутність відокремлюваного префікса. В структуру лінгвістичних даних німецької мови додається таблиця **trent,** яка задає перелік типів відокремлюваних префіксів. А в таблицю **nom** додатково введено поле trnt, де задається тип відокремлюваного префікса для дієслів відповідної групи.

Зв'язки між таблицями даних для німецької мови показані на рис. 4. Зв'язок між таблицями **nom**, **indent** відбувається за

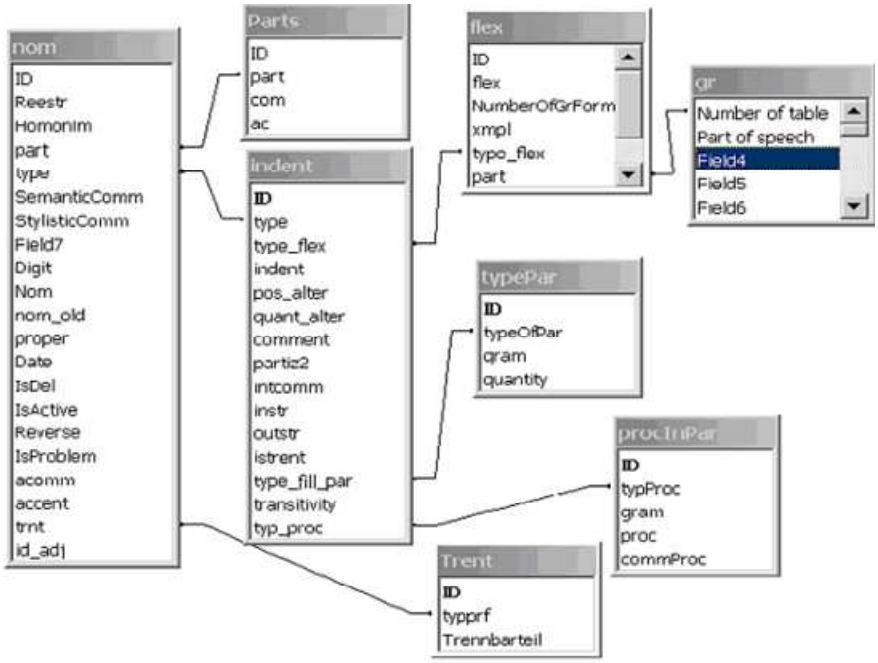

Рис. 4. Схема зв’язків між таблицями ЛБД німецької мови

номером парадигматичного класу (поле type); між таблицями **indent, flex** – за полем номером типового набору квазіфлексій (поле type_flex); між таблицями **nom**, **Parts** – за полем part. Таблиці **indent** та **procIn-Par** пов'язані за полем typProc, а таблиці **trent** і **nom** – за полем typPrf. Полю part таблиці **flex** відповідає поле number of table таблиці **gr**.

Далі наводимо докладний опис полів табл. 2–9. Зазначимо, що призначення аналогічних полів таблиць ЛБД різних мов є однаковим (з тією різницею, що для певної мови якісь поля не використовуються).

Таблиця **nom** проіндексована за полями: ID (unique), Reestr, Homonim, part, type, Digit, Nom, nom_old, proper.

Таблиця 2. Реєстрові одиниці (опис полів таблиці **nom**)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер запису	Лічильник
Reestr	Реєстрове слово	Текстовий
Homonim	Номер омонімії	Числовий
part	Код граматичного класу	Числовий
type	Номер парадигматичного класу	Числовий
SemanticComm	Семантичний коментар	Текстовий
StylisticComm	Стилістичний коментар	Текстовий
Field7	Граматичний коментар	Текстовий
Digit	Реєстрова одиниця у вигляді цифрового коду	Числовий
Nom	Зарезервовано	Числовий
nom_old	Унікальний ідентифікатор слова для створення файлу gram.dic	Числовий
proper	Ознака, чи є слово власною назвою; містить також информацію про властивості прийменників і сполучників	Числовий
Date	Дата останнього редагування слова	Дата/час
IsDel	Ознака, чи є слово видаленим	Логічний
IsActive	Ознака, чи є слово активним	Логічний
Reverse	Зворотний цифровий код реєстрового слова (для сортування в інверсному порядку)	Числовий
IsProblem	Ознака, чи є слово проблемним	Логічний
acomm	Робочий коментар для внутрішнього використання	Текстовий
accent	Номер класу наголосів	Числовий
trnt	Тип відокремлюваного префікса (для дієслів); відповідає номеру в таблиці Trent; trnt = 0, якщо немає відокремлюваного префіксу	Числовий
Id_adj	Індекс для відтворення зв'язку між ступенями порівняння, утвореними від однієї основної форми ад'єктива	Текстовий

Таблиця 3. Параметри парадигматичних класів (опис полів таблиці **indent**)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер запису	Лічильник
type	Номер парадигматичного класу	Числовий
type_flex	Номер типового набору флексій	Числовий
indent	Позиція (від кінця слова) - скільки символів потрібно відрізати для одержання квазіоснови (кількість символів квазіфлексії)	Числовий
pos_alter	Номер позиції від кінця слова, починаючи з якої виділяється підрядок, у якому відбувається зміна (чергування)	Числовий
quant_alter	Кількість букв, що входять у підрядок, який підлягає заміні на послідовність символів, записаних у полі outstr	Числовий
comment	Поле для коментарів	Текстовий
intcomm	Поле для коментарів	Текстовий
outstr	Послідовність символів, на яку заміняється instr	Текстовий
istrent	Клас з відокремлюваною приставкою	Логічний
transitivity	Перехідність	Текстовий
type_fill_par	Тип заповнювання парадигми	Числовий
typ_proc	Номер типового набору процедур (утворення аналітичних форм)	Числовий
partiz2	Спосіб утворення Partizip-2	Числовий

Таблиця **indent** проіндексована за полями: ID (unique), type, type_flex, comment, transitivity, type_fill_par, typ_proc.

Таблиця 4. Набори квазіфлексій (опис полів таблиці **flex**)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер запису	Лічильник
flex	Квазіфлексія	Текстовий
NumberOfGrForm	Номер граматичного значення (див. Таблицю **gr**)	Числовий
xmpl	Приклад слова	Текстовий
type_ flex	Номер парадигматичного класу (номер типового набору квазіфлексій)	Числовий
part	Код класу слів (з таблиці **gr**))	Числовий
comm_fl	Коментар щодо форми (типу: рідко, застаріле, тощо)	Текстовий

Таблиця **flex** проіндексована за полями: ID (unique), NumberOfGrForm, part, type_flex.

Таблиця 5. Опис полів таблиці **gr**

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер запису	Лічильник
number of table	Код класу слів	Числовий
part of speech	Назва класу слів	Текстовий
Field4, Field5, ..., Field29	Граматичні значення	Текстовий

Таблиця **Parts** проіндексована за полями: id (unique), com.

Таблиця 6. Опис полів таблиці **Parts** (граматичні класи)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер запису	Лічильник
part	Номер граматичного класу	Числовий
com	Назва граматичного класу	Текстовий
ac	Додатковий коментар	Текстовий

Таблиця **procInPar** проіндексована за полями*:* id (unique), typProc.

Таблиця 7. Типи процедур утворення аналітичних форм (Таблиця **procInPar**)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер	Лічильник
typProc	Номер типового набору процедур побудови аналітичних форм	Числовий
gram	Номер граматичного значення	Числовий
proc	Тип процедури	Числовий
commProc	Опис процедури	Текстовий

Таблиця 8. Таблиця **Trent** (відокремлювані префікси)

Поле	Призначення (опис)	Тип даних
ID	Унікальний номер	Лічильник
typprf	Тип відокремлюваного префікса (номер)	Числовий
trennbarteil	Відокремлювана частина слова	Текстовий

Таблиця 9. Таблиця **typePar** (Типи заповнення парадигми)

Поле	Призначення (опис)	Тип даних
id	Унікальний номер запису	Лічильник
type_fill_par	Тип заповнювання парадигми	Числовий
gram	Номер граматичного значення	Числовий
quantity	Кількість граматичних форм у відповідному грам. значенні	Числовий

Програмний інтерфейс для підготовки та редагування граматичних ЛБД

Інтерфейс Л-системи ЕГС розроблено з використанням елементів керування операційного середовища Windows. Доступ користувача до кожного з модулів Л-системи ЕГС забезпечується спеціальною інтерфейсною програмою.

Головне вікно програми поділено на три зони:

1) функціональна зона;
2) реєстрова зона;
3) зона 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 (рис. 5) – містить пункти “Файл”, “Вигляд”, “Словник”, “Загальний вибір”, “Вибірка” і “Довідка”. Кожен з перелічених пунктів меню містить підменю:

“Файл” – “Вихід”;

“Вигляд” – “Панель інструментів”, “Рядок стану”;

“Словник” – “Російський”, “Англійський”, “Німецький”, “Іспанський”; “Прямий”, “Інверсний”;

“Загальний вибір” – “Всі”, “Всі з вилученими”, “Тільки вилучені”, “Тільки активні”, “Тільки неактивні”, “Вилучені та неактивні”;

“Вибірка” – “Всі”, “Іменник”, “Прикметник”, “Числівник”, “Займенник”, “Дієслово”, “Дієприкметник”, “Незмінювані”, “Омоніми”, “Власні назви”;

“Довідка” – “Допомога”, “Про програму”.

Рис. 5. Загальне меню

Підзона з інструментарієм для виконання основних функцій має вигляд, показаний на рис. 6. Вибір необхідної функції Л-системи здійснюється за допомогою відповідних кнопок. Кнопка “П” – функція “Парадигма” (за умовчанням завжди активна), кнопка “Т” – функція “Транскрипція” (в даній версії цю функцію не реалізовано). Наступні кнопки призначені для виконання таких функцій: “Введення нового слова”: “Копіювання вибраного з реєстру слова”; “Видалення вибраного слова з реєстру”; “Запис в текстовий файл парадигми вибраного слова або вибраної з реєстру групи слів”; “Перехід до режиму редагування парадигматичних класів”.

Рис. 6. Інструментарій для редагування

Вибірка груп слів з реєстру (крім можливостей, передбачених у Загальному меню), може виконуватись за номером парадигматичного класу, а також за довільним запитом на мові SQL. Таку можливість користувачеві надає фрагмент функціональної зони, зображений на рис. 7. Кнопка “ПК.” і текстовий блок (edit box), розташований справа від неї, призначені для виконання запиту на виведення частини реєстру за заданим номером парадигматичного класу. Кнопка “SQL” призначена для виконання SQL-запиту, який записується у текстовому блоці, розташованому справа від кнопки “І”; кнопка “І” призначена для перевірки тексту запиту.

Рис. 7. Інтерфейс для вибірки слів за парадигматичним класом або SQL-запитом

Інтерфейс для пошуку слова складається з текстового блоку (edit box) для введення пошукового слова і кнопки “Пошук” (рис. 8).

Рис. 8. Інтерфейс для пошуку слова

Реєстрова зона (list box) складається власне з реєстру. В стовпчику “ПК” поруч з реєстровим словом наводиться номер парадигматичного класу, до якого це слово належить. Якщо реєстрова одиниця не є словозмінною, номер парадигматичного класу не наводиться (рис. 9).

Реєстр	^	№	П.К.	D
а		1		
а		2		
а		3		
Аарóн		3.	3137	
абá		7	1502	
абажýр		9	1151	
абажýрный		1.	831	
абáз		1.	1151	
абазúн		1.	1010	
абазúнец		3.	1171	
абазúнка		1.	1635	
абазúнский		1.	808	
Абáй	1	3.	3018	

Рис. 9. Фрагмент реєстрової зони

Зона лексикографічної інформації призначена для відображення інформації зі словозміни обраного з реєстру слова (повна словозмінна парадигма), (рис. 10).

Далі наводимо діалогові вікна, які використовуються при реалізації таких функцій словника як введення нового слова, редагування слова з реєстра, копіювання реєстрового слова, редагування, введення і вилучення парадигматичних класів. При виконанні перших трьох із перелічених функцій використовується діалогове вікно, показано на рис. 11.

абажур - существительное мужского рода

падеж	единственное число	множественное число
именительный	абажу́р	абажу́ры
родительный	абажу́ра	абажу́ров
дательный	абажу́ру	абажу́рам
винительный	абажу́р	абажу́ры
творительный	абажу́ром	абажу́рами
предложный	абажу́ре	абажу́рах

м 1а

Рис. 10. Зона лексикографічної інформації

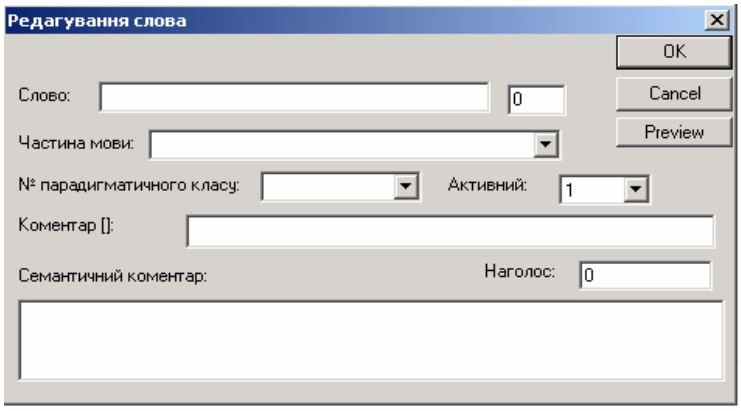

Рис. 11. Вікно редагування слова

Виконання операцій з парадигматичними класами (додавання, редагування, вилучення) відбувається за допомогою вікна, показаного на рис. 12.

Вікно має три зони: функціональну, зону парадигматичних класів (зліва) та зону квазіфлексій (справа). Функціональна зона знаходиться у верхній частині вікна і містить такі елементи керування для виконання функцій з парадигматичними класами: кнопка “Пошук” і розташований поруч справа текстовий блок реалізують пошук парадигматичного класу в списку парадигматичних класів; кнопка “Додати паради-

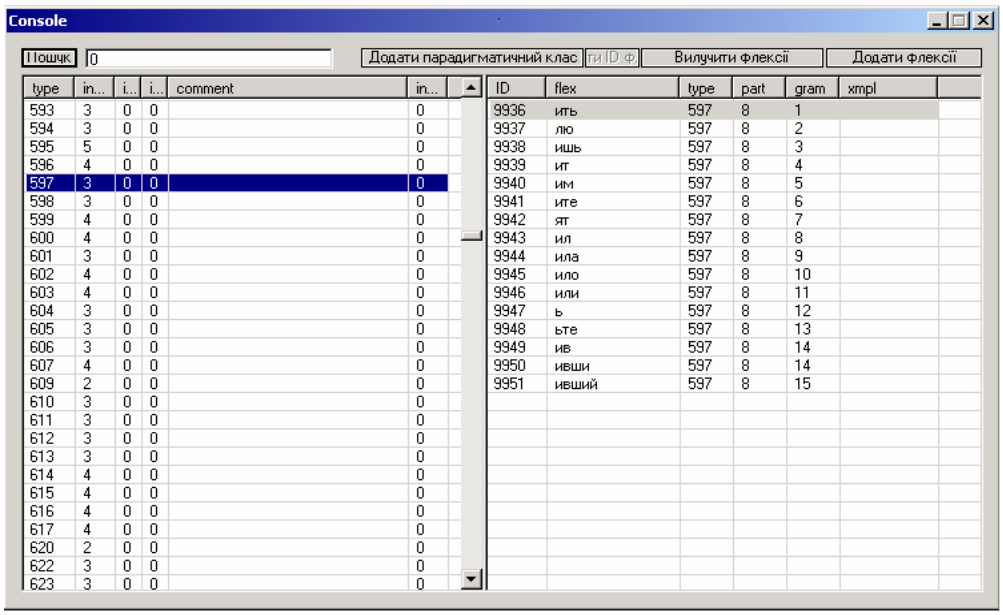

Рис. 12. Діалогове вікно для виконання операцій з парадигматичними класами

гматичний клас" використовується для введення нового парадигматичного класу. При натисненні цієї кнопки активізується діалогове вікно введення нового парадигматичного класу (рис. 13). Кнопки "Вилучити флексії" та "Додати флексії" забезпечують виконання відповідних функцій обраного парадигматичного класу.

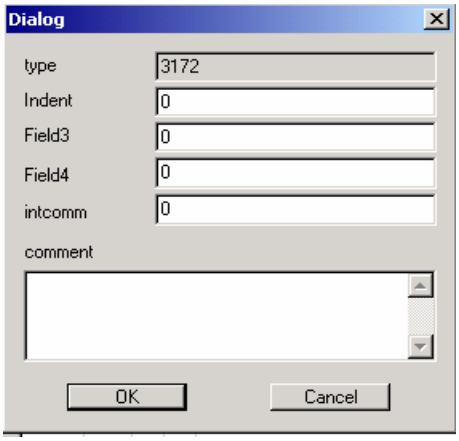

Рис. 13. Діалогове вікно введення нового парадигматичного класу

Редагування і поповнення граматичних ЛБД

Створені граматичні ЛБД функціонують під СУБД Microsoft SQL Server 7.0. Клієнтську програму редагування ЛБД ЕГС розроблено і створено в середовищі Microsoft Visual Studio 6.0. Програма працює під управлінням операційної системи Microsoft Windows 2000 або Microsoft Windows XP. Програма орієнтована на роботу в мережевому середовищі.

Програма реалізує такі функції:

- перегляд реєстру;
- отримання повної словозмінної парадигми обраного з реєстру слова та його основних граматичних характеристик;
- вивід і перегляд частини реєстру (за частиною мови, за номером парадигматичного класу, за довільним запитом (на мові SQL);
- видача всіх граматичних омонімів, власних імен, тощо;
- видача кількісних характеристик щодо наповнення парадигматичних класів, частин мови, омонімів тощо;
- пошук слів у реєстрі;
- побудова прямого або інверсійного словника (встановлення прямого або інверсійного сортування в реєстрі);
- введення нових та редагування вже наявних реєстрових слів, видалення слів із реєстру;
- введення, редагування, видалення парадигматичних класів (задавання їхніх диференційних характеристик; введення і редагування квазіфлексій синтетичних форм та типів процедур утворення аналітичних форм (для аналітичних або аналітико-синтетичних мов);
- запис у файл або вивід на друк виділених фрагментів (наприклад, вивід повної парадигми певного слова; запис у файл частини реєстру тощо);
- побудова словника квазіоснов (для мов флективного типу; словник квазіоснов використовується програмами морфологічного та синтаксичного аналізу).

Робоче вікно програми показано на рис. 14.

Усі програми редагування ЛБД розроблено й створено в середовищі Microsoft Visual Studio 6.0. Програми працюють під управлінням операційної системи Microsoft Windows 2000 або Microsoft Windows XP.

Відзначимо, що створені граматичні словники російської та української[4] мов, а також програмний інструмент пройшли апробацію на значних лексичних масивах (російська мова – близько 170 тис. лексичних одиниць; німецька мова – 60 тис. (іменники, ад'єктиви, дієслова), англійська – близько 20 тис. (іменники, дієслова); для іспанської мови здійснено комп'ютерний експеримент на матеріалі дієслів). Результати експериментальної експлуатації дозволяють вважати, що підхід з єдиних позицій (у рамках пропонованої концептуальної моделі) може бути застосований для широкого (деякого) класу мов із флективною словозміною.

[4] Граматична Л-система української мови та відповідна ЛБД розвинена у працях І.В. Шевченка див., наприклад [20].

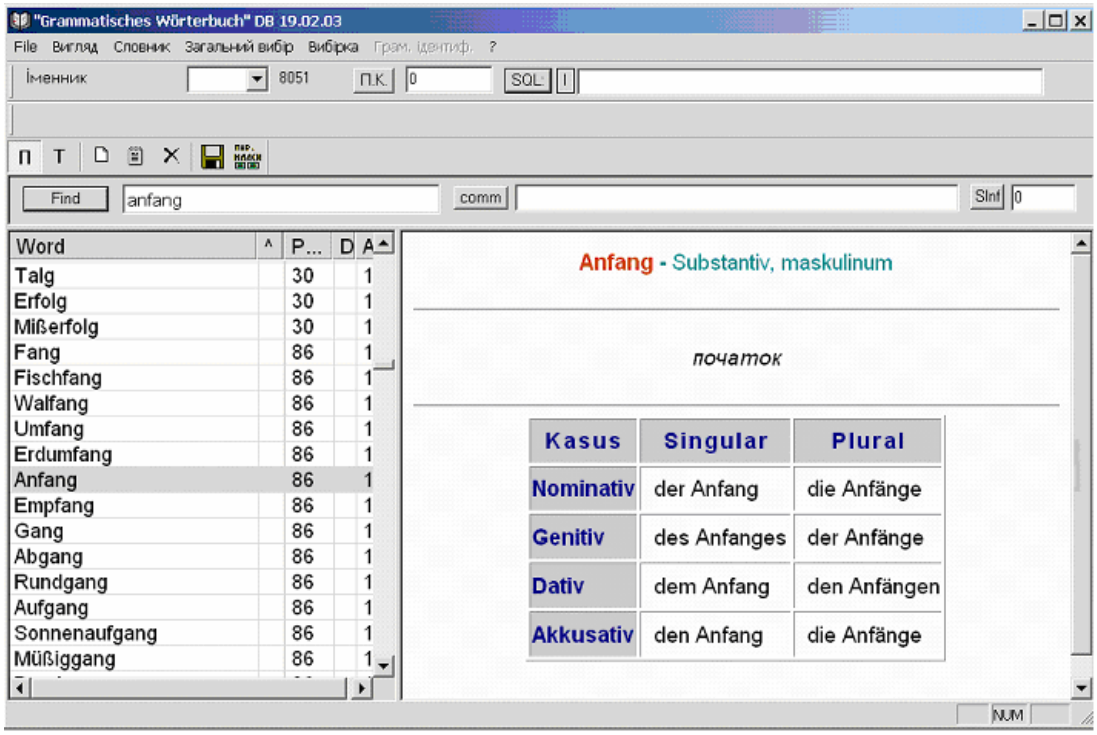

Рис. 14. Робоче вікно програми редагування граматичного словника

Висновки

Розроблені ЛБД лексикографічних систем ЕГС та програмні засоби їх редагування дозволяють ефективно організувати процес створення граматичних словників.

Створені ЛБД можуть успішно використовуватись при дослідженні словозмінних процесів і явищ, зокрема таких, які важко здійснити в “ручному” режимі.

Для такого інструменту, як універсальна граматична Л-система, важливою властивістю є мовна незалежність (у деякому класі мов).

З використанням описаного підходу до моделювання і програмування граматичних лексикографічних систем створені ЛБД для різних мов: спочатку для російської мови, потім для німецької, англійської і, частково, іспанської (дієслова). Отже є підстави вважати, що підхід до моделювання і програмування граматичних лексикографічних систем з єдиних позицій (у рамках розглянутої концептуальної моделі) може бути застосований і для більш широкого класу мов із флективною словозміною.

1. *Широков В.А.* Інформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – 331с.
2. *Широков В.А.* Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії // Мовознавство. – 2002. – № 6. – С. 7 – 48.
3. *Широков В.А., Рабулець О.Г., Костишин О.М.* та ін. Технологічні основи сучасної тлумачної лексикографії // Там само. –2002. – № 6. – С. 49 – 86.
4. *Широков В.А.* Феноменологія лексикографічних систем. – К.: Наук. думка, 2004. – 327 с.
5. *Белоногов Г.Г., Зеленков Ю.Г.* Алгоритм морфологического анализа русских слов // Вопросы информационной теории ипрактики. – М.: ВИНИТИ, 1985. – № 53.
6. *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л.* и др. Лингвистическое обеспечение системе автоматического перевода Этап-2. – М.: Наука, 1989. – 296 с.
7. *Бидер И.Г., Большаков И.А., Еськова Н.А.* Формальная модель русской морфологии. – ИРЯ АН СССР, Проблемная группа по экспериментальной и прикладной лингвистике. – Вып. 112. – М.: 1978. – I – 60 с., II – 59 с.
8. *Хорошилов А.А.* Автоматическая нормализация слов в системах обработки научно-технической информации: Автореф. дис. … канд. техн. наук. – М.; 1987. – С. 18.

9. *Гельбух А.Ф., Сидоров Г.О.* К вопросу об автоматическом морфологическом анализе флективных языков // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. "Диалог'2005" (Звенигород, 1–6 июня, 2005). – http://www.dialog-21.ru/Archive/2005/Gelbukh%20Sidorov/GelbukhA.htm
10. *Лингвистический* энциклопедический словарь / Гл. ред. В.Н. Ярцева. М.: Советская энциклопедия, 1990. – 688 с.
11. *Корпусна* лінгвістика: Монографія / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна, Т.П. Любченко, О.Г. Рабулець, О.О. Сидоренко, Н.М. Сидорчук, І.В. Шевченко, О.О. Шипнівська, К.М. Якименко; Український мовно-інформаційний фонд НАН України. – К.: Довіра, 2005. – 472 с.
12. *Любченко Т.П.* Організація даних та структура електронного граматичного словника німецької мови // Математичні машини і системи: наук. журнал. – 2007. – № 2.
13. *Зализняк А.А.* Грамматический словарь русского языка: Словоизменение. – М.: Русский язык, 1978. – 878 с. (4-е изд., испр. и доп.: 2003).
14. *Грязнухина Т.А., Любченко Т.П., Рабулец А.Г.* Электронная версия грамматического словаря русского языка (А.А. Зализняк) как инструмент автоматического морфологического анализа русского текста // Докл. науч. конф. «Корпусная лингвистика и лингвистические базы данных». – Санкт-Петербург, март 2002. – С. 63–70.
15. *Wahrig G.* Deutsches Wörterbuch. Wissen Media Verlag GmbH, Gütersloch/ München 2002 (vormals Bertelsmann Lexikon Verlag GmbH). – 1451 s.
16. *Helbig G., Buscha J.* Deutsche Grammatik. VEB Verlag Enzyklopädie Leipzig, 1979. – 629 s.
17. *Садиков А.В., Нарумов Б.П.* Испанско-русский словарь современного употребления. – М.: Русский язык, 2001. – 752 с.
18. *Collins Cobuild*. Collins Birmingham University International Language Database. Student's Grammar. Practice Material by Dave Willis. – Harper Collins Publishers 1991. – 263 p.
19. *English* Conjugation: System and Functioning (reference-book) / Под ред. В.И. Перебейнос – К.: КНЛУ, 2003. – 3,3 МВ. – Електронне видання.
20. *Шевченко І.В*. Моделі та алгоритмічно-програмне забезпечення лексикографічних систем : Автореф. дис. ... канд. техн. наук. – К., 2000. – С. 20.

Отримано 17.01.2007

Про автора:

Любченко Тетяна Петрівна,
молодший науковий співробітник
Українського мовно-інформаційного
фонду НАН України.

Місце роботи автора:

Український мовно-інформаційний фонд НАН України,
01601, Київ, вул. Володимирська, 54.
Тел.: (044) 525 8165
E-mail: ltp@i.com.ua