УДК 658.012.011.56

Т.П. ЛЮБЧЕНКО

ОРГАНІЗАЦІЯ ДАНИХ ТА СТРУКТУРА ЕЛЕКТРОННОГО ГРАМАТИЧНОГО СЛОВНИКА НІМЕЦЬКОЇ МОВИ

Abstract: *Problems of development of the grammar German dictionary structure, his functionalities and some aspects of constructing the user interface are examined. Dictionary is intended for using in an information and reference system, as well as for application in the language processing systems (morpho-syntactic analysis and text synthesis).*
Key words: *digital grammatical dictionary, German, data structure.*

Анотація: *У статті обговорюються питання розробки структури електронного граматичного словника німецької мови та його функціональні можливості, а також деякі аспекти побудови користувацького інтерфейсу. Словник призначається для застосування в контурах систем автоматичної обробки текстової інформації (в алгоритмах морфолого-синтаксичного аналізу та синтезу текста), а також для використання його в інформаційно-довідковій системі.*
Ключові слова: *електронний граматичний словник, німецька мова, структура даних.*

Аннотация: *Обсуждаются вопросы разработки структуры электронного грамматического словаря немецкого языка и его функциональные возможности, а также некоторые аспекты построения пользовательского интерфейса. Словарь предназначается для применения в контурах систем автоматической обработки текстовой информации (в алгоритмах морфолого-синтаксического анализа и синтеза текста), а также для использования в информационно-справочной системе.*
Ключевые слова: *электронный грамматический словарь, немецкий язык, структура данных.*

1. Вступ

Електронний граматичний словник (ЕГС) німецької мови розробляється як частина інтегрованої лексикографічної системи Українського мовно-інформаційного фонду НАНУ (УМІФ НАНУ). В УМІФ НАНУ електронні граматичні словники розробляються для мов, які передбачається включити до системи багатомовного машинного перекладу (МП): української, російської, англійської, німецької, іспанської, французької та турецької мов. Зазначені словники орієнтовано на письмові варіанти мов. ЕГС призначені, насамперед, для використання їх в алгоритмах морфологічного (морфолого-синтаксичного) аналізу в системі МП (на етапах морфологічної розмітки тексту, лематизації та синтезу). Окрім цього, передбачено надання користувачеві можливості доступу до словника як до довідково-інформаційної системи (пошук слів, надання інформації відносно словозміни певних реєстрових одиниць). В основу розробки ЕГС покладено теорію лексикографічних систем [1–4].

В даній статті ми зупинимося на проблемах створення граматичного словника німецької мови.

2. Джерела лінгвістичної інформації

При створенні ЕГС німецької мови було використано відповідні граматики і словники [5–7] та [9–12]. Як основне джерело лінгвістичної інформації зі словозміни німецької мови використовувався Словник Герхарда Вариға (G. Wahrig, Deutsches Wörterbuch [6] (у подальшому – DW), у якому подано найбільш повну інформацію про словозмінну систему німецької мови. Наведену в Словнику DW класифікацію німецької лексики взято за основу.

3. Принципи моделювання словозміни німецької лексики

При побудові морфологічної моделі німецької мови виходимо з принципів, які були застосовані при розгляді таких флективних мов, як російська (та українська) [3, С. 218–225; 14]. Відмінними в

моделях словозміни різних мов є лише набори словозмінних параметрів, притаманних мові, що розглядається.

Побудова концептуальної моделі словозміни німецької лексики починається зі встановлення та формалізації тих лінгвістичних критеріїв, за якими множина усіх слів розбивається на певні підмножини, взаємний перетин яких є пустою множиною (порожнім), і ці підмножини є такі, що всередині кожної з них словозміна відбувається за єдиним алгоритмом. Такі підмножини слів (підмножини із такими властивостями) називатимемо *парадигматичними класами*.

(Під парадигматичним класом ми розуміємо групу лексем, словозмінна парадигма яких характеризується однаковою кількістю граматичних форм, усередині якої словозміна відбувається за тим самим (єдиним) правилом. Для німецької мови, яка є мовою аналітико-синтетичного типу, це означає, що, по-перше, слова, які належать до одного парадигматичного класу, мають однакові флексії у відповідних граматичних значеннях та однаковий характер чергування в основі і, по-друге, відповідні аналітичні форми будуються за однаковими моделями їх утворення).

Парадигматичні класи визначаються нами всередині кожного класу слів (це викликано тим, що словозмінні процеси для слів різних класів визначаються своїми, притаманними саме для цього класу, словозмінними параметрами).

Словниковий склад поділяється нами на такі класи: іменники, ад'єктиви, дієслова, артиклі, займенники та незмінювані. Деякі з цих класів за своїми класифікуючими ознаками розділяються ще на декілька підкласів.

Наведемо систему ознак (граматичних категорій), що визначають словозміну зазначених класів слів, та значення, які ці ознаки (категорії) можуть приймати.

Іменники

• Рід* (Genus)* – чоловічий (m), жіночий (f), середній (n), множинний (pl).

• Тип відмінювання (Deklination)* – сильний (stark), слабкий (schwach), мішаний (gemischt), ад'єктивний (adjektivisch).

• Відмінок (Kasus) – називний (N), родовий (G), давальний (D), знахідний (A).

• Число (Numerus) – однина (S), множина (P).

Дієслова

• Тип дієвідміни* – слабкий (schwach), сильний (stark).

• Перехідність (Trasitivität) – перехідний (vt), неперехідний (vi) (впливає на наявність / відсутність пасивних форм у словозмінній парадигмі дієслова).

• Спосіб утворення дієприкметника Partizip2 – (1) – для дієслів з невідокремлюваним префіксом (verloren), (2) – для дієслів з відокремлюваним префіксом (eingerichtet, angerufen), (3) – стандартний спосіб, суть якого полягає у застосуванні такої схеми: префікс „ge-" + основа+ флексія (gefragt, gefahren).

• Стан (Genus) – активний (Aktiv), пасивний (Passiv).

• Спосіб (Modus) – дійсний (Indikativ), наказовий (Imperativ), умовний (Konjunktiv).

• Час (Tempus) – теперішній (Präsens), минулий (Präteritum (Imperfekt), минулий (Perfekt), давноминулий (Plusquamperfekt), майбутній (Futurum1), майбутній (Futurum2).

• Число (Numerus) – однина (Singular), множина (Plural).

• Особа (Person) – 1., 2., 3.

• Допоміжне дієслово для утворення аналітичних форм (Hilfsverb): sein, haben.

Ад'єктиви

• Ступінь порівняння* (Komparation, Steigerung) – основна форма (Positiv), вищий ступінь (Komparativ), найвищий ступінь (Superlativ).

• Рід (Genus) – чоловічий (m), жіночий (f), середній (n), множинний (pl).

• Число (Numerus) – однина (S), множина (P).

• Відмінок (Kasus) – називний (N), родовий (G), давальний (D), знахідний (A).

• Вид артикля – означений, неозначений, без артикля.

• Застосування – prädikativ, attributiv, adverbial.

Артиклі

• Вид артикля* – означений (bestimmt), неозначений (unbestimmt).

• Рід (Genus) – чоловічий (m), жіночий (f), середній (n), множинний (pl).

• Відмінок (Kasus) – називний (N), родовий (G), давальний (D), знахідний (A).

• Число (Numerus) – однина (S), множина (P).

Займенники

• Тип займенника* – особові (Personalpronomen), зворотні (Reflexivpronomen), невизначені (Indefinitpronomen), вказівні (Demonstrativpronomen), питальні (Interrogativpronomen), присвійні (Possesivpronomen), відносні (Relativpronomen).

• Відмінок (Kasus) – називний (N), родовий (G), давальний (D), знахідний (A).

• Рід (Genus) – чоловічий (m), жіночий (f), середній (n).

• Число (Numerus) – однина (S), множина (P).

• Особа (Person) – 1., 2., 3.

Зірочкою позначені граматичні категорії, які є класифікаційними для конкретного класу слів (тобто клас слів, у якого є категорія, позначена значком «*», за значеннями цієї категорії розподіляється на декілька підкласів, а саме на стільки, скільки значень ця категорія може приймати. Наприклад, рід у іменників є класифікаційною ознакою: це означає, що клас іменників за ознакою роду розподіляється на 4 підкласи). Всі інші ознаки є словозмінними.

Нехай W – множина відмінюваних слів німецької мови. Розіб'ємо множину W на підмножини (класи слів), що взаємно не перетинаються:

$$W = \bigcup_{i=1}^{5} W_i , \qquad (1)$$

де W_1 – іменники, W_2 – дієслова, W_3 – ад'єктиви, W_4 – артиклі, W_5 – займенники.

Омонімію будемо вважати знятою, а омоніми промаркованими, так що $W_{j_1} \cap W_{j_2} = \varnothing$ при $j_1 \neq j_2, \quad j_1, j_2 = 1,2,...,5.$

Як вже було зазначено, деякі класи слів за своїми класифікаційними ознаками поділяються на підкласи. Так, розподіл іменників на підкласи відбувається у два етапи: за типом відмінювання (слабкий, сильний, мішаний та ад'єктивний) та за значенням граматичної категорії *рід*.

$W_1 = \bigcup_{j=1}^{4} W_1^{\,j}$, де $W_1^{\,1}$ – іменники слабкого типу відмінювання, $W_1^{\,2}$ – іменники сильного типу, $W_1^{\,3}$ – іменники мішаного типу відмінювання, $W_1^{\,4}$ – іменники ад'єктивного типу відмінювання.

$W_1^{\,1} = \bigcup_{j=1}^{4} W_1^{\,1j}$, де $W_1^{\,11}$ – іменники слабкого типу відмінювання чоловічого роду, $W_1^{\,12}$ – іменники слабкого типу жіночого роду, $W_1^{\,13}$ – іменники слабкого типу середнього роду ([1]), $W_1^{\,14}$ – множинні іменники слабкого типу відмінювання.

$W_1^{\,2} = \bigcup_{j=1}^{4} W_1^{\,2j}$, де $W_1^{\,21}$ – іменники сильного типу відмінювання чоловічого роду, $W_1^{\,22}$ – іменники сильного типу жіночого роду, $W_1^{\,23}$ – іменники сильного типу середнього роду, $W_1^{\,24}$ – множинні іменники сильного типу відмінювання.

$W_1^{\,3} = \bigcup_{j=1}^{4} W_1^{\,3j}$, де $W_1^{\,31}$ – іменники чоловічого роду мішаного типу відмінювання, $W_1^{\,32}$ – іменники жіночого роду мішаного типу[2], $W_1^{\,33}$ – іменники середнього роду мішаного типу, $W_1^{\,34}$ – множинні іменники мішаного типу відмінювання.

$W_1^{\,4} = \bigcup_{j=1}^{4} W_1^{\,4j}$, де $W_1^{\,41}$ – іменники чоловічого роду ад'єктивного типу відмінювання, $W_1^{\,42}$ – іменники жіночого роду ад'єктивного типу, $W_1^{\,43}$ – іменники середнього роду ад'єктивного типу, $W_1^{\,44}$ – множинні іменники ад'єктивного типу відмінювання.

Отже, клас іменників

$$W_1 = W_1^{11} \cup W_1^{12} \cup W_1^{13} \cup W_1^{14} \cup W_1^{21} \cup W_1^{22} \cup W_1^{23} \cup W_1^{24} \cup W_1^{31} \cup W_1^{32} \cup W_1^{33} \cup W_1^{34} \cup W_1^{41} \cup W_1^{42} \cup W_1^{43} \cup W_1^{44}.$$

Підкласи $W_1^{\,jk}$ $(j,k = 1,2,3,4)$ будемо називати граматичними класами іменників і позначатимемо їх у подальшому P_i, $i = 1,2,...,16$: $P_i \equiv W_1^{\,jk}$, $j,k = 1,2,3,4$.

Клас дієслів розбивається за типом дієвідміни на слабкі (schwach) та сильні (stark):

$W_2 = W_2^{\,1} \cup W_2^{\,2}$, де $W_2^{\,1}$ – дієслова слабкої дієвідміни, $W_2^{\,2}$ – дієслова сильної дієвідміни.

Підкласи $W_2^{\,j}$ $(j = 1,2)$ будемо називати граматичними класами дієслів. Позначимо $P_i \equiv W_2^{\,j}$, $i = 17,18;\ j = 1,2$.

Клас ад'єктивів розбивається за ступенем порівняння:

[1] Множина $W_1^{\,13}$ є порожньою, оскільки в німецькій мові немає іменників середнього роду слабкого типу відмінювання.
[2] Множина $W_1^{\,32}$ є порожньою, оскільки в німецькій мові немає іменників жіночого роду мішаного типу відмінювання.

$W_3 = W_3^1 \cup W_3^2 \cup W_3^3$, де W_3^1 – ад'єктиви позитивного ступеня (основної форми), W_3^2 – компаративні ад'єктиви (вищого ступеня порівняння), W_3^3 – суперлативні ад'єктиви (найвищого ступеня). Підкласи $W_3^j (i = 1,2,3)$ будемо називати граматичними класами ад'єктивів. Позначимо $P_i \equiv W_3^j, \quad i = 19,20,21; \; j = 1,2,3.$

Займенники розподіляються на підкласи таким чином:

$W_4 = W_4^1 \cup W_4^2 \cup W_4^3 \cup W_4^4 \cup W_4^5 \cup W_4^6 \cup W_4^7$, де W_4^1 – особові займенники (Personalpronomen), W_4^2 – зворотні (Reflexivpronomen), W_4^3 – невизначені (Indefinitpronomen), W_4^4 – вказівні (Demonstrativpronomen), W_4^5 – питальні (Interrogativpronomen), W_4^6 – присвійні (Possesivpronomen), W_4^7 – відносні займенники (Relativpronomen). Підкласи W_4^j $(j = 1,2,...,7)$ будемо називати граматичними класами займенників. Позначимо $P_i \equiv W_4^j, \quad i = 22,23,...,28; \; j = 1,2,...,7$.

В результаті для кожного слова $x \in W$ однозначно визначається приналежність до певного граматичного класу P_j, $j = 1,2,...28$. (Маємо 28 граматичних словозмінних класів: $W = \bigcup_{i=1}^{28} P_i$. $P_{j_1} \cap P_{j_2} = \varnothing, \; j_1 \neq j_2, \; j_1, j_2 = 1,2,...,28$.)

Усередині граматичних класів виділяємо парадигматичні класи.

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

$$x = c(x) * f(x), \tag{2}$$

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

Змінна та незмінна складові можуть мати як нульову довжину, так і представляти собою всю лексему. Наприклад, у парадигмах із суплетивними формами незмінна частина дорівнює нулю, а змінна частина представлена всіма словоформами (*bin*, *bist*, *ist, sind*, *seid*,...). У парадигмах незмінних слів, навпаки, нулю дорівнює змінна частина.

Повна словозмінна парадигма $[x]$ слова x, що належить до граматичного класу W_i, представляється у вигляді

$$\pi(x) = c(x) * \{ f_i(x) \}, \tag{3}$$

де $f_i(x), i = 0,1,2,...,n_i$ – змінні частини слова (квазіфлексії) у відповідних граматичних значеннях; причому в деяких із них може існувати більше однієї словоформи. Для означення даного факту введемо параметр кратності граматичної форми $\nu(w_i(x))$, який задається цілим числом, рівним кількості можливих форм лексеми x у i-тому граматичному значенні. У загальному випадку

$$f_i(x) = \bigcup_{l=0}^{\nu(w_i(x))} f_{il}\ , \tag{4}$$

$l = l(i) = 0,1,2,...$ – індекс кількості словоформ у i-тому граматичному значенні;

$f_0(x)$ – 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_i – кількість граматичних значень у граматичному класі W_i.

Покладемо

$$\begin{aligned} F &= \bigcup_{x \in W} (\{f_0(x)\}, \{f_{1l}(x)\}, ..., \{f_{n_i l}(x)\}) \equiv \\ &\equiv \{f_{jl}^{\ 1}, f_{jl}^{\ 2}, ..., f_{jl}^{\ N_i}\},\ j = 0,1,2,...,n_i,\ l = l(w_j) = 0,1,2,... \end{aligned} \tag{5}$$

Тоді

$$F = \bigcup_{k=1}^{N_i} [F]^k \text{ , де } [F]^k = \{f^k\} = \{f_{jl}^{\ k}, j = 0,1,...,n_i\}. N_i = N(W_i),\ l = l(w_i). \tag{6}$$

Таким чином, кожна множина $[F]^k$ складається з квазіфлексій слів, які мають у всіх своїх граматичних формах $w_1, w_2, ..., w_{n_i}$ однакові змінні складові.

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

$$\forall x^1, x^2 \in P_i\ x^1 \pi_i x^2 : x^1 = c(x^1) * f^k, x^2 = c(x^2) * f^k, f^k \in [F]^k. \tag{7}$$

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

Фактор-множина P_i / π_i є множиною парадигматичних класів граматичного класу P_i. Очевидно, що різні словозмінні парадигматичні класи не перетинаються. Отже P_i є об'єднанням 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 $c(x)$. Слова з одного класу еквівалентності, визначеного в такий спосіб, мають і однакові правила словозміни.

Таким чином, для кожного з граматичних класів 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.

На сьогодні створено парадигматичну класифікацію іменників, ад'єктивів та дієслів німецької мови.

Визначено 456 парадигматичних класів, серед них 85 парадигматичних класів іменників, 6 класів ад'єктивів та 365 класів дієслів.

4. Структура даних ЕГС німецької мови

При розробці моделі даних німецької мови (з метою уніфікації представлення їх у лексикографічних базах даних (ЛБД) різних мов) було використано підхід, аналогічний до підходу, який застосовується нами для мов флективного типу [3, с. 225–231], [13]. При цьому моделі конкретної мови доповнюються необхідними даними, що враховують особливості словозмінної системи цієї мови.

Німецька мова характеризується такими 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; чергування в основі; 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, відповідних полів.

Структура даних електронного граматичного словника репрезентується реляційною моделлю, яка включає такі таблиці:

• таблиця **nom**, яка подає реєстрові одиниці *Reestr* разом з кодом граматичного класу *part* та номером парадигматичного класу (поле *type*);

• таблиця **indent**, яка задає параметри та характеристики, що є однаковими для певного парадигматичного класу;

• таблиця квазіфлексій **flex**, де для кожної граматичної форми (поле *NumbOfGrForm*) кожного парадигматичного класу (поле *type*) задано квазіфлексії *flex*;

• таблиця **Parts** граматичних класів та їх кодів;

• таблиця **gr** словозмінних типів;

• таблиця **procInPar** типових наборів *typProc* процедур утворення аналітичних форм;

• таблиця **trent**, яка задає перелік відокремлюваних префіксів та їх кодів *typPrf;*

• **typePar** (Типи заповнення парадигми).

Нижче наводимо докладний опис полів усіх таблиць.

Таблиця 1. Реєстрові одиниці (Опис полів таблиці **nom**)

Поле	Призначення (опис)	Тип даних
id	Унікальний номер запису	Лічильник
reestr	Реєстрове слово	Текстовий
field2	Номер омонімії	Числовий
part	Код граматичного класу	Числовий
type	Номер парадигматичного класу	Числовий
field5	Семантичний коментар	Текстовий
field6	Стилістичний коментар	Текстовий
field7	Переклад	Текстовий
digit	Реєстрова одиниця у вигляді цифрового коду	Числовий
nom	Зарезервовано	Числовий
nom_old	Унікальний ідентифікатор слова для створення файлу gram.dic	Числовий
own	Ознака, чи є слово власною назвою; містить також информацію про властивості прийменників і союзів	Числовий

Продовження табл. 1

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 префікса	Числовий

Таблиця **nom** проіндексована за полями: *id* (unique), *reestr*, *field2*, *part*, *type*, *digit*, *nom*, *nom_old*, *own*.

Таблиця 2. 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		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
typproc	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*, *typproc*.

Таблиця 3. Набори квазіфлексій (Опис полів таблиці **flex**)

Поле	Призначення (опис)	Тип даних
id	Унікальний номер запису	Лічильник
flex	Квазіфлексія	Текстовий
field2	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), *field2*, *part*, *type_flex*.

Таблиця 4. 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 5. 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 **Parts** проіндексована за полями: *id* (unique), *com*.

Таблиця 6. 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 **procInPar** проіндексована за полями: *id* (unique), *typProc*.

Таблиця 7. 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 8. 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

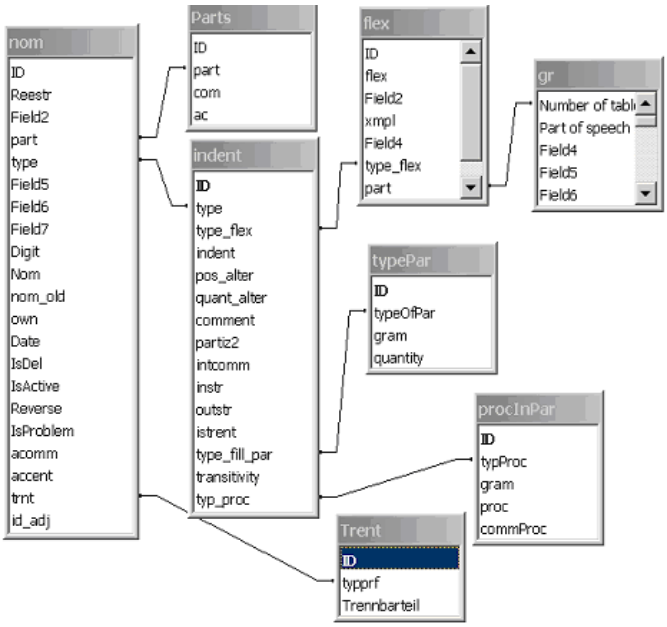

Рис. 1. Схема зв'язків між таблицями ЛБД німецької мови

Зв'язки між таблицями показані на рис. 1. Зв'язок між таблицями **nom**, **indent** відбувається за 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 ***procInPar*** пов'язані за полем *typProc*, а таблиці **trent** і **nom** – за полем *typPrf*. Поле *part* таблиці **flex** відповідає полю *number of table* таблиці **gr**.

5. Програмний інтерфейс для підготовки та редагування граматичної ЛБД

Інтерфейс лексикографічної системи ЕГС (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: Функціональна зона; Реєстрова зона; 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. 2) 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" – "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".

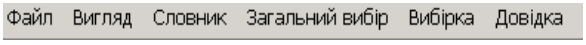

Рис. 2. Загальне меню

Підзона з інструментарієм для виконання основних функцій має вигляд, наведений на рис.3. 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. 3. 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. 4. Кнопка "П.К." 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 "I"; кнопка "I" призначена для перевірки тексту запиту.

Рис. 4. 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. 5).

Рис. 5. Інтерфейс для пошуку слова

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

Word	^	old	P....	D	A
Andenken		0	16		1
Andrang		0	30		1
Androhung		0	1		1
Anekdote		0	2		1
Anerbieten		0	16		1
Anerkennung		0	1		1
Anfahrt		0	1		1
Anfall		0	86		1
Anfang		0	86		1
Anfangszeit		0	1		1
Anfänger		0	40		1
Anfechtung		0	1		1
Anflug		0	31		1
Anforderung		0	1		1
Anfrage		0	2		1

Рис. 6. Фрагмент реєстрової зони

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

Рис. 7. Зона лексикографічної інформації

6. Супровід (редагування і поповнення) граматичної ЛБД

Граматична ЛБД функціонує під СУБД Microsoft SQL Server 7.0. Клієнтську програму супроводу (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 відносно 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 їх диференційних характеристик; введення та редагування квазіфлексій – для флективних мов, типів процедур утворення аналітичних форм для мов аналітичних);

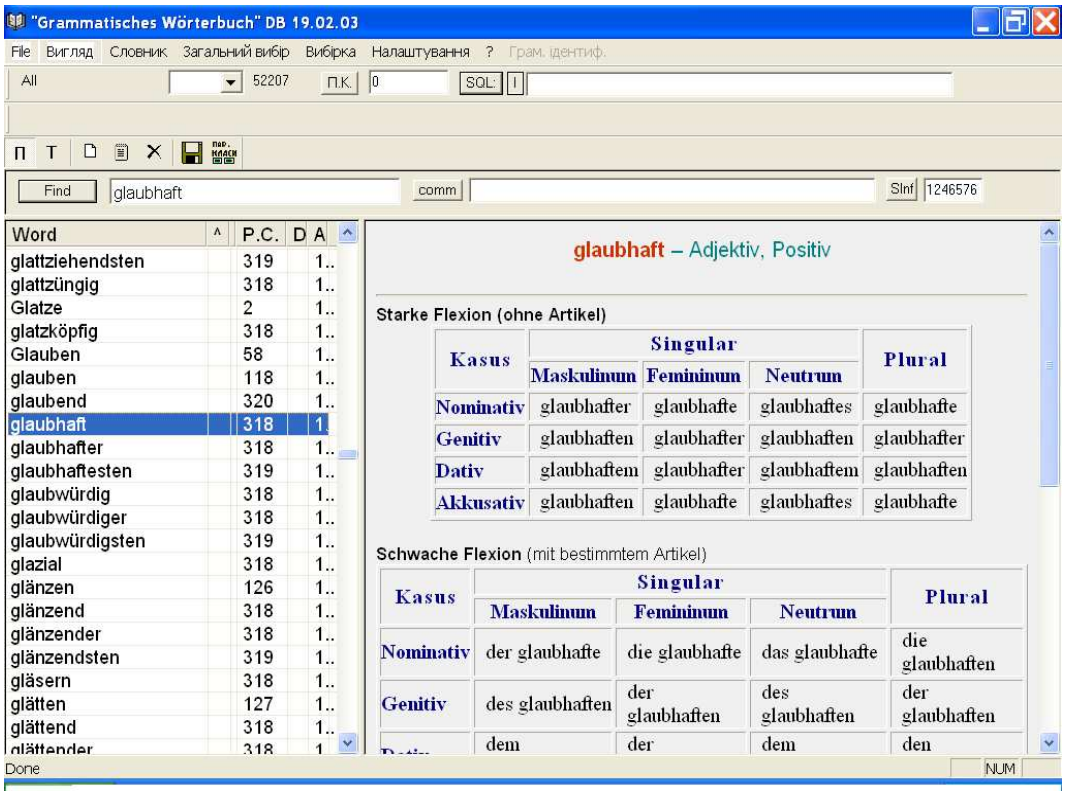

Рис. 8. Робоче вікно програми редагування німецького граматичного словника

• 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 зображено на рис. 8.

7. Висновки

У роботі описано принципи комп'ютерного моделювання словозміни німецької лексики, запропоновано формальне визначення поняття парадигматичного класу, розроблено класифікацію лексики німецької мови за парадигматичними класами. Роботу виконано на масиві німецької лексики обсягом понад 52 тис. лексем.

Розроблена структура бази даних ЕГС та програмні засоби редагування даних словника дозволяють ефективно організувати процес створення граматичного словника.

Створена граматична лексикографічна база даних німецької мови може успішно використовуватись при дослідженні словозмінних процесів і явищ, зокрема, таких, які важко було б провести в "ручному" режимі.

Передбачено створення граматичних ЛБД для інших мов, залучених до системи МП, яка розробляється в УМІФ НАНУ (англійської, іспанської, російської). Розглянуті у статті принципи моделювання системи словозміни німецької мови знаходять застосування й для інших мов. Звичайно, кожна мова має свої особливості, урахування яких спонукає до відповідних змін у структурі даних, а також розробки нових алгоритмів і програм. Паралельно зі створенням ЛБД для згаданих мов буде виконуватися розробка алгоритмів та програмних модулів морфологічного (морфолого-синтаксичного) аналізу текстів, написаних відповідними мовами.

СПИСОК ЛІТЕРАТУРИ

1. Широков В.А. Інформаційна теорія лексикографічних систем. – Київ: Довіра, 1998. – 331 с.
2. Широков В.А. Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії // Мовознавство. – 2002. – № 6. – С. 7–48.
3. Широков В.А. та ін. Корпусна лінгвістика: Монографія / Широков В.А., Бугаков О.В., Грязнухіна Т.О., Любченко Т.П., Рабулець О.Г., Сидоренко О.О., Сидорчук Н.М., Шевченко І.В., Шипнівська О.О., Якименко К.М.; Український мовно-інформаційний фонд НАН України. – К.: Довіра, 2005. – 472 с.
4. Інтегрована лексикографічна система «Словники України» / Широков В.А., Шевченко І.В., Рабулець О.Г., Костишин О.М., Якименко К.М. – Київ, 2004 (електронне видання, версія 1.04).
5. Русско-немецкий словарь (основной): Ок. 53000 слов / Под ред. Лейна. – Киев: Русский язык, 1989. – 736 с.
6. Wahrig G. Deutsches Wörterbuch. Wissen Media Verlag GmbH, Gütersloch. – München, 2002 (vormals Bertelsmann Lexikon Verlag GmbH). – 1451p.
7. Helbig G., Buscha J. Deutsche Grammatik. – VEB Verlag Enzyklopädie Leipzig, 1979. – 629 p.
8. Любченко Т.П. Синтез словоформ німецьких іменників у системі машинного перекладу // Вісник лінгвістичного університету. – Київ, 2002. – Т. 5, № 2. – С. 145–154.
9. Lezius W. Morphologiesystem MORPHY / http://www.lezius.de/wolfgang/morphy/papers.html.
10. http://www.canoo.net.
11. http://www-psycho.uni-padeborn.de/lezius/.
12. http://wortschatz-uni.leipzid.de/.
13. Любченко Т.П. Технология создания системы автоматической парадигматической классификации русского языка // Искусственный интеллект. – 2002. Материалы Международной научно-технической конференции. – Т. 2. – Таганрог: Изд-во ТРТУ. – 2002. – С. 19–21.
14. Любченко Т.П. Морфологічна модель словозміни флективної мови та електронний граматичний словник // Біоніка інтелекту: Науково-технічний журнал. – 2006. – № 1 (64). – С. 72–77.