

Таким чином, кожна множина $[F]^k$ складається з квазіфлексій слів, які мають у всіх своїх граматичних формах $w_1, w_2, \dots, w_{n(i)}$ парадигматичного типу t_i однакові змінні складові.

Оскільки $[F]^k$ побудовано у такий спосіб, що до них увійшли унікальні набори квазіфлексій, тобто $[F]^i \neq [F]^j$ при $i \neq j$ ($i, j = 1, 2, \dots, N_i$), то для кожного парадигматичного типу t_i можна побудувати відношення π_i на декартовому добутку $t_i \times t_i$, яке визначається так:

$$\forall x^1, x^2 \in t_i \quad \pi_i x^1, x^2: x^1 = c(x^1) * f^k, x^2 = c(x^2) * f^k, f^k \in [F]^k. \quad (2.28)$$

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

Фактор-множина t_i/π_i є множиною парадигматичних класів парадигматичного типу t_i .

Очевидно, що 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 парадигматичних типів t_i будується розбиття на множині слів, що не перетинаються, і які є парадигматичними класами, всередині кожного з яких діють єдині правила словозміни. Для мов флективного типу це передбачає однаковість флексій граматичних форм та збіг характеру чергування в основі.¹¹

Граматичні значення в структурі даних ЕГС представлено двосимвольними кодами: перший символ позначає граматичний клас слів, другий — граматичне значення словозмінної форми.

Як було показано, належність двох лексем до одного і того ж СПК означає, що будь-яка словоформа однієї лексеми відрізняється від словоформи іншої лексеми з тими ж граматичними параметрами лише своєю незмінною частиною. Використовуючи набір алгоритмів, здійснюється граматична ідентифікація лексеми, тобто розбиття реестрового ряду словника за СПК. За поширеністю на лексичному масиві СПК дуже різні.

Наприклад, іменниковий СПК (системний № 1592 в Електронному граматичному словнику УМІФу) включає лише дві лексеми:

¹¹ Для мов аналітико-синтетичних та синтетико-аналітичних до цих вимог додається вимога однаковості моделей утворення аналітичних форм.

“Великдень” і “переддень”. Дієслівний СПК=263, що включає, зокрема, дієслово “наполягти” 14 лексем (усі — префіксальні утворення від кореня —ляг-). У той же час найпоширеніший прикметниковий СПК №2302 (прикметники твердої групи без особливостей в узгодженні) включає 18238 лексем. Числівниковий СПК №2408 (кількісні числівники на -дцять) включає 17 лексем, а займенниковий клас №1164 (присвійні займенники м’якої групи) обмежений трьома словами: *мій, свій, твій*.

Загалом в ЕГС УМІФу виокремлено 2053 словозмінні класи, з яких приблизно половина припадає на дієслівні класи.

Створений в УМІФі електронний граматичний словник побудовано на основі граматичної бази даних української мови, яка узагальнює матеріали новітніх академічних словників. ЕГС охоплює такі показники: вихідна форма слова; приналежність до омонімічного ряду; частина мови; рід для іменникового; вид дієслова; парадигматичний (словозмінний) і акцентуаційний клас; повна словозмінна парадигма з позначенням наголосів для кожної словоформи; додаткові характеристики слова: обмеженість вживання (діалектизм, просторічне, архаїзм), стисле тлумачення значення для омонімів та рідкісних слів, джерело, рік включення слова в реєстр і т.п.

ЕГС експлікує в явному вигляді всі словоформи для тих слів, які належать до змінюваних частин мови. Реєстр словника розташований в його лівій частині. Кожному слову ставиться у відповідність його СПК і словозмінна парадигма, що формується автоматично.

Кожна реєстрова одиниця в граматичному словнику споряджена наголосом, причому у вихідній словоформі наголоси проставлено згідно з Орфографічним словником української мови. Було розроблено акцентуаційну класифікацію, яка ставить у відповідність слову набір схеми наголошення у похідних словоформах. Для лексем нульового класу зсув наголосів в похідних словоформах не відбувається. Таких лексем для української мови близько 90% у загальному масиві. Проте для приблизно 25 тисяч (на масиві 232 тисяч українських реєстрових одиниць) різноманітні зсуви позицій наголосу мають місце. Наприклад, акцентуаційний клас №1 охоплює дієслова на -увати з наголосом на суфіксі —увати (напр. абеткувати). Вони мають цілком специфічну схему наголошення. В той же час дієслова з наголосом на -ювати, якщо вони мають у парадигмі пасивний дієприкметник (малювати — мальований), то по-

трапляють в інший клас завдяки зсуву в цьому граматичному значенні на 1 символ ліворуч (для класу 1 цей зсув складає -2). Загалом виділено 691 акцентуаційний клас. Спеціальна програма автоматично перевіряє правильність наголосів в словоформах ЕГС (навність наголосів на приголосних, дефісах і тому подібне).

ЕГС дозволяє пошук слова за його маскою, вибірку всіх слів певної частини мови або парадигматичного класу, омонімів, незмінних слів, власних імен. Реєстр може виводитися на екран за абеткою або в інверсійному порядку. Застосування SQL-запитів до бази даних граматичного словника дозволяє формувати і більш складні вибірки.

Підсистема ведення і актуалізації граматичного словника дозволяє переносити лексему в інший парадигматичний або акцентуаційний клас у разі неточності в парадигмі або зміни правопису, змінювати існуючі і створювати нові парадигматичні класи. Є можливість змінювати статус лексем в ЕГС, фіксуючи їх як активні, неактивні, вилучені. Загалом українська частина ЕГС нараховує 553293 лексеми, включно з створеними за словозмінними моделями та не зафіксованими в словниках (такі слова позначені як неактивні). Активних лексем в ЕГС всього нараховується 232750.

2.2. Структура бази даних ЕГС

Структура даних електронного граматичного словника флективної мови репрезентується реляційною моделлю. Дані представлені такими таблицями:

- таблицею **nom** реєстрових одиниць *Reestr*;
- таблицею квазіфлексій **flex**, в якій для кожної граматичної форми (поле *NumbOfGrForm*) кожного парадигматичного класу (поле *type*) задані квазіфлексії *flex*;
- таблицею **indent**, що задає параметри та характеристики, які є однаковими для кожного з парадигматичних класів;
- таблицею **Parts** лексико-граматичних класів та їхніх кодів;
- таблицею **gr** словозмінних типів;
- таблицею наголосів “**accent**”.

Перелічені таблиці з повним описом усіх полів представляємо нижче.

Фрагмент таблиці реєстрових одиниць української мови лот

ID	Rcestr	Field2 part	Field5	Field6	Field7	Digit	Nom	nom_old	own	Date	IsDel	IsActive	Reverse	IsProblem	asomni	ascent
1	a	1	72	0		01	1	1	0	07.04.2005	HI	TAK	01	HI		0
2	a	2	73	0		01	2	2	2	19.08.2002	HI	TAK	01	HI		0
3	a	3	74	0		01	3	3	0	11.11.2002	HI	TAK	01	HI	2005/02 SUM	0
5	абажур	0	8	1607		010201092 421	5	5	0	15.11.2002	HI	TAK	2124090 10201	HI		0
6	абажурний	11	2302			010201092 421181114	6	6	0	19.08.2002	HI	TAK	1411182 1240901 0201	HI		0
7	абажурник	0	8	1776		010201092 421281115	7	7	0	15.11.2002	HI	TAK	1511282 1240901 0201	HI		0
8	абаз	8	1607			010201101 118	8	8	0	19.08.2002	HI	TAK	10010201	HI		0
9	абазин	7	1766	(пред- ставник народ- ності)		010201101 118	9	9	0	19.08.2002	HI	TAK	1811100 10201	HI		0
10	абазинць	7	1540			010201101 118072731	10	10	0	19.08.2002	HI	TAK	3127071 8111001 0201	HI		0

- ID** — унікальний ідентифікатор запису;
Reestr — реєстрова одиниця;
Field2 — номер омонімії (якщо 0 — омонімії немає);
part — частина мови (з таблиці Parts);
type — номер парадигматичного класу;
Field5 — семантичний коментар;
Field6 — стилістичний коментар;
Digit — реєстрова одиниця у вигляді цифрового коду (використовується для сортування);
Nom — зарезервовано;
nom_old — унікальний ідентифікатор слова для створення файла gram.dic;
own — ознака, чи є слово власною назвою; також зберігає властивості сполучників та прийменників;
Date — дата останньої зміни слова;
IsDel — ознака, чи є слово вилученим;
IsActive — ознака, чи є слово активним;
Reverse — реєстрові одиниці у вигляді зворотнього цифрового коду (для сортування та створення інверсного словника);
IsProblem — ознака, чи є слово проблемним;
acomm — робочий коментар для внутрішнього використання;
accent — номер класу наголосів;

Таблиця проіндексована за полями: *ID* (unique), *Reestr*, *Field2*, *part*, *type*, *Digit*, *Nom*, *nom_old*, *own*, *acomm*.

Таблиця 2.2

Таблиця квазіфлексій *flex*:

ID	flex	Field2	xmpl	Field4	type	part
37571	іль	1			2132	1
37572	олі	2			2132	1
37573	олі	3			2132	1
37574	іль	4			2132	1
37575	іллю	5			2132	1
37576	олі	6			2132	1
37577	оле	7			2132	1

ID — унікальний ідентифікатор запису;
flex — квазіфлексія;
Field2 — номер граматичного значення;
xmpl — приклад слова;
Field4 — зарезервовано;
type — номер парадигматичного класу;
part — частина мови;

Таблиця проіндексована за полями: *ID* (unique), *Field2*, *type*, *part*.

Таблиця 2.3.

Таблиця *indent*:

<i>ID</i>	<i>type</i>	<i>indent</i>	<i>Field3</i>	<i>Field4</i>	<i>comment</i>	<i>intcomm</i>
2130	2130	3	2	1		0
2131	2131	3	2	1		0
2132	2132	3	0	0		0
2133	2133	4	0	0		0
2134	2134	1	0	0		0
2135	2135	1	0	0		0

ID — унікальний ідентифікатор запису;

type — номер парадигматичного класу;

indent — кількість літер, які потрібно відрізати від кінця слова, щоб залишилась квазіоснова (наприклад: “дивовижність”, *type* — 2130; квазіоснова — “дивовижні”, квазіфлексія — “сть”);

Field3 — номер літери від початку квазіфлексії, з якої починається незмінна частина квазіфлексії (наприклад: “дивовижність”, *type* — 2130; квазіфлексія — “сть”, незмінна частина починається з “т”);

Field4 — кількість літер, що відрізаються у квазіфлексії (починаючи з *Field3*), щоб отримати незмінну частину квазіфлексії (наприклад: “дивовижність”, *type* — 2130; квазіфлексія — “сть”, незмінна частина починається з другої літери — “т”, незмінна частина квазіфлексії — “т”);

comment — граматичний коментар до парадигматичного класу;

intcomm — зарезервовано;

Таблиця проіндексована за полями: *ID* (unique), *type*, *comment*, *intcomm*.

Таблиця наголосів *accent*:

ID	indent1	indent2	indent3	indent4	accent_type	part	gram	xmpl
779	0	-2			37		15	загнистися
780	0	-2			37		16	
781	0	255			37		19	
782	0	-1			37		20	
783	0	-1			37		21	
784	0	-1			37		22	
785	0	255			37		26	

ID — унікальний ідентифікатор запису;

indent1 — кількість позицій, на які потрібно змістити перший наголос слова з тої позиції, яку він займає в початковій формі (якщо дорівнює 0, то наголос залишається на місці; якщо дорівнює 255, то наголос зникає);

indent2 — те ж саме для другого наголосу (якщо в початковій формі другого наголосу не було, то його позиція в даній формі відраховується від позиції першого наголосу в початковій);

indent3 — те ж саме для третього наголосу (якщо в початковій формі третього наголосу не було, то його позиція в даній формі відраховується від позиції другого наголосу в початковій; якщо не було і другого наголосу, то від позиції першого);

indent4 — те ж саме для четвертого наголосу (якщо в початковій формі четвертого наголосу не було, то його позиція в даній формі відраховується від позиції третього наголосу в початковій; якщо не було і третього наголосу, то від позиції другого);

accent_type — номер класу наголосу;

part — частина мови;

gram — номер граматичного значення;

xmpl — приклад слова;

Таблиця проіндексована за полями: *accent_type*, *gram*.

Таблиця граматичних значень *gr* містить такі поля:

Number of table — код частини мови;

Part of speech — назва частини мови;

Field4, Field5, ..., Field29 — значення граматичних категорій (граматичне значення).

Таблиця граматичних класів **Parts** містить такі поля:

ID — унікальний ідентифікатор запису, який є кодом частини мови;

part — скорочена назва частини мови;

com — повна назва частини мови;

ac — додатковий коментар.

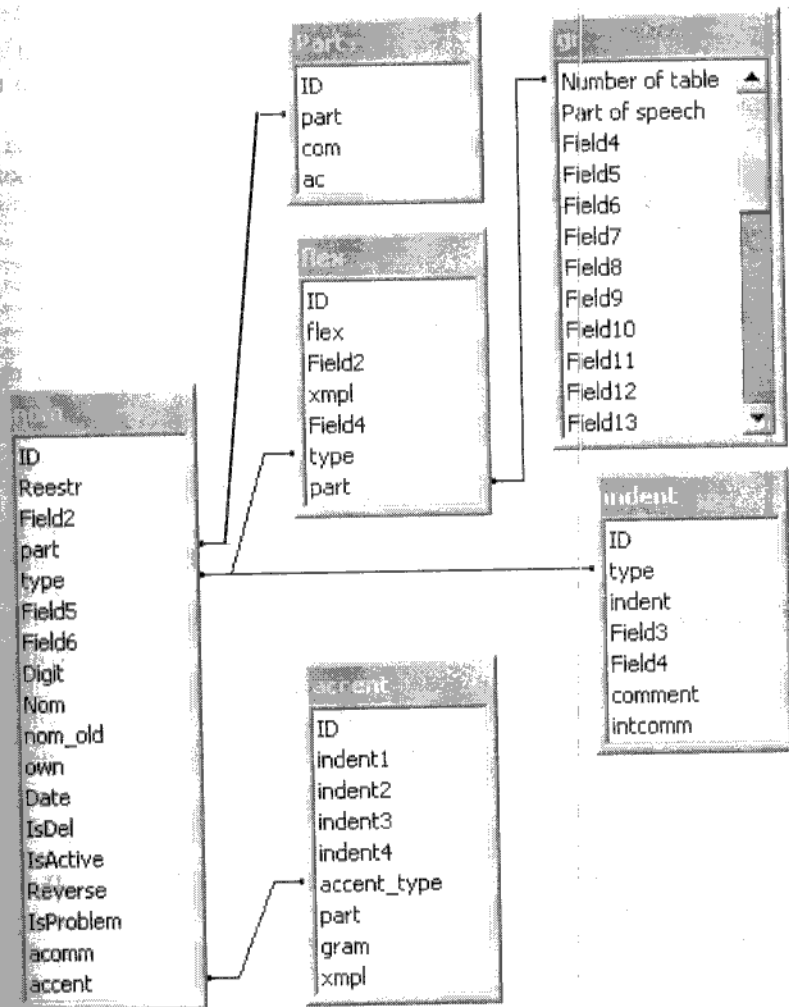


Рис. 2.1. Схема зв'язків між таблицями даних для української мови

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

Між частинами мови з таблиць **gr** та **Parts** існує відповідність: значенням поля *Number of table* таблиці **gr** відповідають певні значення поля *ID* таблиці **Parts**.

Також до граматичної ЛБД входить таблиця **TYP_REFL**, яка складається з одного поля *type* і фактично є переліком дієслівних парадигматичних класів, які є рефлексивними.

Ця схема є спрощеною: вона не відображає непрямі зв'язки, які існують від полів *Field2* таблиці *flex* та *gram* таблиці *accent* до номерів полів *Field4*—*Field29* таблиці *gr*.

Структура даних ЕГС російської мови в основному подібна до описаної вище структури даних української мови. Відмінності полягають у кодуванні частин мови та граматичних значень, а також у відповідності між частинами мови з таблиць **gr** та **rParts**.

2.3. Зовнішній рівень представлення ЕГС

Для роботи з граматичною ЛБД була створена клієнтська програма редагування граматичної ЛБД. Ця програма дозволяє виконувати такі функції:

- перегляд реєстру граматичного словника, повної парадигми та транскрипції кожного слова;
- перехід до російського словника (при цьому програма працює з таблицями такої самої структури, як описано вище, але вони мають в своїй назві префікс “r”);
- сортування в прямому або інверсному режимі;
- фільтрація реєстру: за частинами мови, за парадигматичними класами, за омонімами, за власними назвами, за складеними словами, за активними/неактивними та вилученими/невилученими реєстровими одиницями, за довільним запитом;
- пошук слів в реєстрі;
- додавання, вилучення та редагування реєстрових слів;
- додавання, вилучення та редагування парадигматичних класів та флексій в парадигматичних класах;
- редагування частин, що відрізаються;
- запис у файл реєстрових одиниць або статей;
- перевірка наголосів;
- створення граматичного файлу *gram.dic*.

Головне вікно програми має такий вигляд:

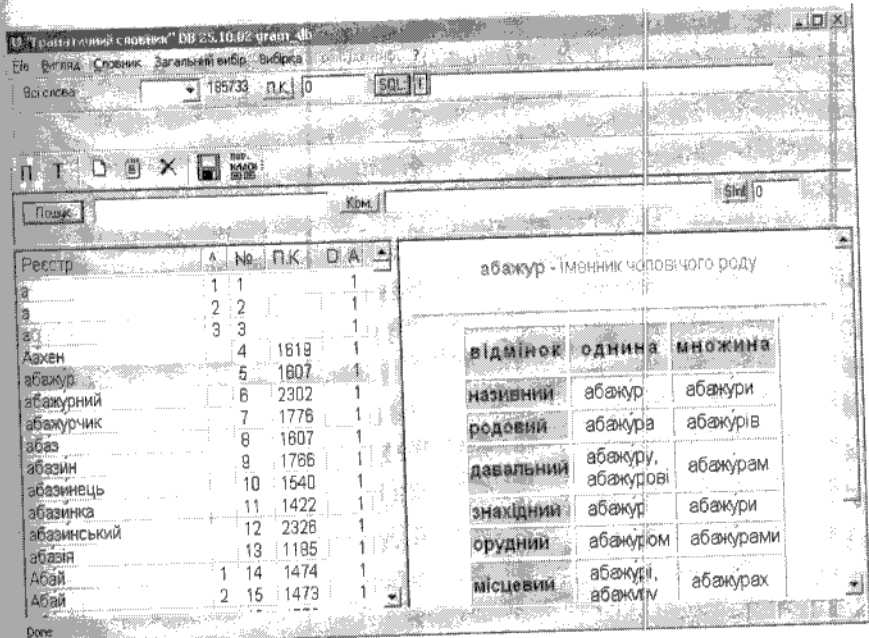


Рис. 2.2. Робоче вікно програми для роботи з ЛБД української мови

Розглянемо більш детально останню функцію. Структура описаної вище граматичної ЛБД є досить зручною для формування ЛБД та її редагування, але для програм, що потребують швидкого опрацювання парадигми слів вона не є оптимальним варіантом збереження граматичної інформації з погляду швидкодії та дискового простору, що займає ЛБД. Тому було розроблено оптимізовану ЛБД, яку реалізує файл типу *gram.dic*, а також бібліотеку функцій для роботи з файлом такої структури (*wpar01.dll*)¹². Така бібліотека може бути динамічно під'єднана до зовнішньої програми ("Словники України", ЛК тощо).

Файл типу *gram.dic* складається з чотирьох компонентів, записаних в ньому послідовно у вигляді:

- розмір компоненти 1 (ціле, 4 байти);
- компоненти 1;
- розмір компоненти 2 (ціле, 4 байти);
- компоненти 4;

¹² Автор — Рабулець О.Г.

На проміжному етапі ці компоненти записуються у файли під назвами *base.wrd* (1), *base.prm* (2), *flex.wrd* (3) та *flex.prm* (4), а потім інтегруються в єдиний файл. Далі розглянемо їх внутрішню структуру:

1. Квазіоснови:

- 1.1. Кількість квазіоснов (ціле, 4 байта).
 - 1.2. Номер версії (Unicode, 6 байтів); якщо для нумерації квазіоснов використовується *nom_old*, то в номері версії замість крапки записується двокрапка (наприклад, "1:0").
 - 1.3. Службова інформація, яка визначається режимом фільтрації реєстру (Unicode, 10 байтів):
 - 1.3.1. "_all" — всі слова, крім вилучених;
 - 1.3.2. "_adel" — взагалі всі слова;
 - 1.3.3. "_del" — тільки вилучені слова;
 - 1.3.4. "_act" — активні, не вилучені слова;
 - 1.3.5. "_nact" — не активні, не вилучені слова;
 - 1.3.6. "_dnact" — не активні, вилучені слова;
 - 1.3.7. "____" — інше.
 - 1.4. Дата створення файлу (у вигляді "ддммгг", Unicode, 12 байтів).
 - 1.5. Квазіоснови (Unicode), що закінчуються на \0; записуються з поля *Reestr* таблиці *nom*.
 - 1.6. Незмінні частини квазіфлексій (Unicode), що закінчуються на \0; формуються з поля *Reestr* таблиці *nom* з використанням даних таблиці *indent*.
- Далі 1.5 та 1.6 повторюються до кінця компоненти.

2. Параметри квазіоснов:

- 2.1. Реєстровий ідентифікатор (ціле, 4 байта); записується з поля *nom_old* таблиці *nom*.
- 2.2. Номер лексичної (?) омонімії (ціле, 1 байт); записується з поля *Field2* таблиці *nom*.
- 2.3. Частина мови (ціле, 1 байт); записується з поля *part* таблиці *nom*.
- 2.4. Номер парадигматичного класу (ціле, 2 байта); записується з поля *type* таблиці *nom*.
- 2.5. Додаткова інформація (ціле, 1 байт):
 - 2.5.1. для активних слів: у сьомий біт записується 1 (x1xxxxxx) (визначається режимом фільтрації реєстру);
 - 2.5.2. для дієслів: якщо двохвидовий, то 1 (xxxxxxx1) (визначається частиною мови *part* з таблиці *nom*); якщо

рефлексивний, то у восьмий біт записується 1 (1xxxxxxx) (визначається таблицею *TYP_REFL*);

2.5.3. для власних назв: антропонім — 2, топонім — 3, присвійний — 4, аббревіатура — 5; записується з поля *own* таблиці *nom*;

2.5.4. для сполучників: підрядний — 1, сурядний — 2, сурядно-підрядний — 3; записується з поля *own* таблиці *nom*;

2.5.5. для прийменників: керує знахідним відмінком — 1 біт (1), керує давальним відмінком — 2 біт (2), керує місцевим відмінком — 3 біт (4), керує родовим відмінком — 4 біт (8), керує орудним відмінком — 5 біт (16); записується з поля *own* таблиці *nom*.

2.6. Інформація з поля *Nom* таблиці *nom* (ціле, 4 байта).

3. Квазіфлексії:

3.1. Кількість парадигматичних класів (ціле, 4 байт); визначається динамічно.

3.2. Службова інформація (аналогічно 1.2—1.4, 28 байт).

3.3. Кількість флексій в класі (ціле, 1 байт); визначається динамічно.

3.4. Частина мови (ціле, 1 байт); записується з поля *part* таблиці *flex*.

3.5. Флексії (Unicode), що закінчуються на \0 — стільки, скільки граматичних значень має парадигматичний клас; записуються з поля *flex* таблиці *flex*, при цьому символи “*”, “^”, “%”, “\$”, “&” та “@” вилучаються.

Далі 3.3—3.5 повторюються стільки раз, скільки є всього парадигматичних класів.

4. Параметри квазіфлексій:

4.1. Номер граматичного значення для частини мови (ціле, 1 байт); записується з поля *Field2* таблиці *flex*.

4.2. Якщо флексія закінчується на “*”, то в перший біт записується 1 (xxxxxxx1); якщо флексія закінчується на “^” (не відображати), то в другий біт записується 1 (xxxxxx1x); якщо флексія закінчується на “%” (по), то в третій біт записується 1 (xxxxx1xx); якщо флексія закінчується на “\$” (на), то в четвертий біт записується 1 (xxxx1xxx); якщо флексія закінчується на “&” (по), то в п’ятий біт записується 1 (xxx1xxxx); якщо флексія закінчується на “@” (до), то в шостий біт записується 1 (xx1xxxxx).

Крім *gram.dic*, створюється також індексний файл *act.ind*. В цей файл записується кількість активних слів (ціле, 4 байта), а також перелік їх реєстрових ідентифікаторів (з поля *nom_old* таблиці *nom*, ціле, 4 байта). Неактивними можуть бути деякі рідковживані, діалектні та інші слова, які не входять до головного реєстру при побудові словника, але мають враховуватись при лематизації та інших процедурах аналізу тексту з використанням *gram.dic*. Файл *act.ind* доцільно використовувати при створенні *gram.dic* в режимі фільтрації реєстру 1.3.1 або 1.3.2.

Опис принципів побудови алгоритмів морфологічного аналізу текстових одиниць, синтезу парадигми конкретного слова та лематизації текстових словоформ з використанням електронного граматичного словника російської мови можна знайти в матеріалах міжнародної конференції “Корпусная лингвистика и лингвистические базы данных 2002”.¹³ Такий же підхід використовується і для української мови.

Модуль морфологічного аналізу працює зі структурно розміченим текстом, в якому вже визначені і роставлені теги для: початку і кінця речення, іншомовних слів, цифр, рубрикаторів, дат, електронних адрес, власних імен тощо. Програма морфологічного аналізу працює за реченнями. В результаті роботи програми МА кожному неомонімічному слову в аналізованому реченні тексту приписується дво(х)символьний код: <частина мови, граматичне значення>. Омонімічні словоформи отримують ланцюжки таких кодів, що відображують їх граматичну омонімію. Для зняття омонімії залучаються алгоритми контекстного аналізу.¹⁴

Покажемо роботу програми морфологічної розметки тексту на конкретному прикладі МА російського тексту.

Головне вікно програми МА наведено на рис. 2.3. Сервіс програми надає можливість вибору мови (російської або української). (Тим самим визначається/ задається, який саме граматичний словник буде використовуватись програмою МА). Програма працює в одному з трьох режимів:

¹³ Т.А.Грязнухина, Т.П. Любченко, А.Г. Рабулец. Электронная версия грамматического словаря русского языка (А.А.Зализняк) как инструмент автоматического морфологического анализа русского текста. / Доклады научной конференции “Корпусная лингвистика и лингвистические базы данных”. — СПб.: Изд-во С.-Петербург. ун-та, 2002, с. 63-70.

¹⁴ Алгоритми контекстного аналізу розроблені Т.О.Грязнухіною.

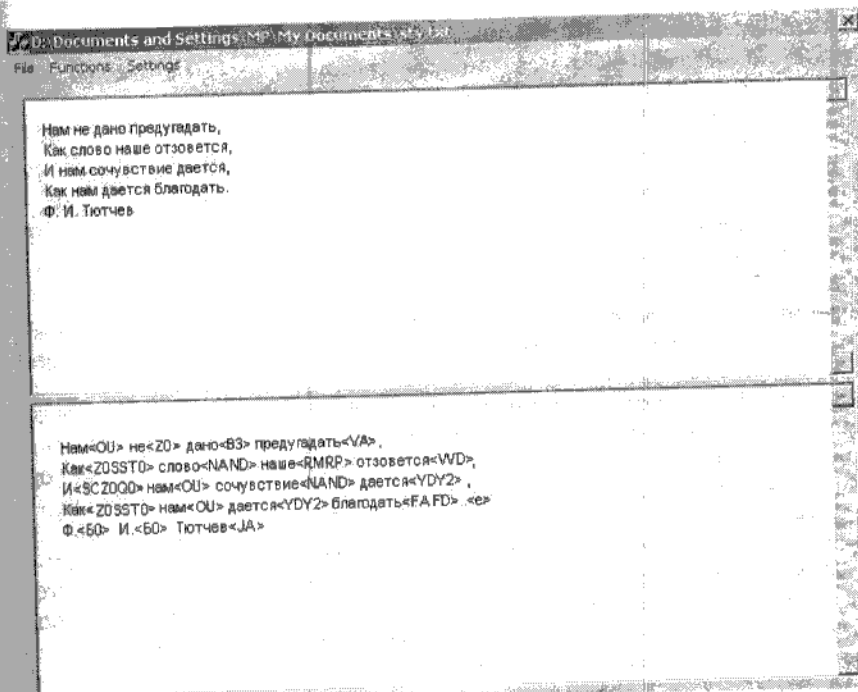


Рис 2.3. Робоче вікно програми МА.

- морфологічний аналіз файлу без завантаження його тексту в робоче вікно. В результаті роботи програми МА формується файл `mrf_*.txt` із проаналізованим і морфологічно розміченим текстом;
- морфологічний аналіз файлу з виводом результату в вікні для перегляду відразу ж після закінчення роботи програми МА;
- морфологічний аналіз тексту — текст для аналізу вміщується в верхній частині робочого вікна (текст у робоче вікно або завантажується з файлу, або вводиться безпосередньо у це вікно); результат МА виводиться для перегляду в нижнє вікно. (На рис. 2.3. показано який вигляд має робоче вікно при роботі програми МА у цьому режимі).

Результатом роботи програми МА російського тексту, наведеного у верхній зоні робочого вікна (рис. 2.3.), є такий текст (текст, який записується в окремий файл):

Нам<OU> = {мы: местоимение, дательный падеж}
 не<Z0> = {не: частица}
 дано<B3> = {данный: причастие, с.р., краткая форма}
 предугадать<VA> = {предугадать: глагол совершенного вида, инфинитив},
 Как<Z0SST0> = {как: частица; союз подчинительный; наречие}
 слово<NAND> = {слово: существительное с.р., ед.число, именительный падеж;
 существительное с.р., ед.ч., винительный падеж}
 наше<RMRP> = {наш: местоимение-прилагательное, с.р., ед.ч., именит. падеж; местоимение-прилагательное, с.р., ед.ч., винит. падеж}
 отзовется<WD> = {отозваться: глагол совершенного вида, будущее время, 3-лицо, ед.ч.},
 И<SCZ0Q0> = {и: союз сочинительный; частица; междометие}
 нам<OU>= {мы: местоимение, дательный падеж}
 сочувствие<NAND>= {сочувствие: существительное с.р., ед.число, именительный падеж; существительное с.р., ед.ч., винительный падеж}
 дается<YDY2>= {даваться: глагол несовершенного вида, возвратный, настоящее время, 3 л., ед. ч.; давать: глагол несовершенного вида, страд. форма на —ся, настоящее время, 3-е лицо, ед. число},
 Как<Z0SST0>= {как: частица; союз подчинительный; наречие}
 нам<OU>= {мы: местоимение, дательный падеж}
 дается<YDY2> = { даваться: глагол несовершенного вида, возвратный, настоящее время, 3 л., ед. ч.; давать: глагол несовершенного вида, страд. форма на —ся, настоящее время, 3-е лицо, ед. число}
 благодать<FAFD>= {благодать: существительное ж.р., ед.ч., именит.падеж; существительное ж.р., ед.ч., винит.падеж} .<e>
 Ф. <Б0>= {инициал}
 И. <Б0>= {инициал}
 Тютчев<JA>= {Тютчев: имя собственное, м.р., ед. число, именительный падеж}

Створені граматичні ЛБД української та російської мов функціонують під СКБД Microsoft SQL Server 7.0. Програми розроблені в середовищі Microsoft Visual Studio 6.0. і працюють під управлінням операційної системи Microsoft Windows 2000 або Microsoft Windows XP.

Значення автоматичного морфологічного аналізу, синтезу і лематизації для прикладних лінгвістичних (і не тільки) досліджень переоцінити важко. Морфологічний модуль потрібен в орфографічних коректорах, в системах оптичного розпізнавання символів, в системах машинного перекладу, в системах штучного інтелекту.

Особливо важливим морфологічний аналіз виявляється для мов флективних, до яких належать українська і російська мови. Кількість різних словоформ в цих мовах така велика, що зберігати граматичну інформацію у вигляді словників словоформ вбачається недоцільним (якби й розвиток обчислювальної техніки і спонукав би до цього). Важливим компонентом систем, які обслуговують корпуси текстів, є лематизатор (який вирішує завдання одержання для кожної довільної словоформи її початкової форми, леми). Виконання лематизації не тільки спрощує побудову пошукових запитів, зменшує “шум” при побудові конкордансів, але й дозволяє отримати для корпусу реєстр (рос. “словник”) словникового типу. Таким чином, у користувача з’являється можливість згенерувати масив слів із прикладами, тобто фактично отримати матеріал для словника, у якому вже будуть позначені словникові статті, що складатимуться з “лексичного входу”, граматичного опису і корпусу прикладів.

Розділ 3

ЛЕКСИКОГРАФІЧНА СИСТЕМА СЛОВНИКА УКРАЇНСЬКОЇ МОВИ

3.1. Словник української мови в 11-ти томах (1970—1980 рр.)

Приклад, який буде розглянуто у цьому розділі — побудова лексикографічної системи тлумачного Словника української мови (СУМ) і на її основі створення лексикографічної бази даних (ЛБД) та комп'ютерної технології укладання тлумачних словників. Урок цієї праці полягає в тому, що застосування теорії лексикографічних систем дозволило здійснити так званий парсинг (конверсію тексту словника в ЛБД) в автоматичному режимі для дуже складного лексикографічного об'єкта, яким є СУМ, — нам невідомі реальні приклади парсингу словників такого великого обсягу та складності — і на цій базі побудувати високоефективну комп'ютерну технологію укладання тлумачних словників. Але докладніше. Фундаментальне академічне видання “Словник української мови” в 11-ти томах¹ заслужено вважається найвищим досягненням нашої національної лексикографії². Він є нормативною і довідковою лексикографічною працею тлумачного типу, що відповідає основним вимогам словникарства і охоплює українську лексику від часів І. Котляревського до 70-х років ХХ століття (див. відгуки³). Базуючись на великому і різноманітному з погляду походження та функціонування лексико-фразеологічному матеріалі, СУМ містить у своєму реєстрі понад 134 тис. слів.

¹ Словник української мови / Л.К. Білодід (гол. ред) та ін. — Т. 1-11. — К.: Наукова думка. 1970-1980.

² Серед перших спроб тлумачної лексикографії в українському мовознавстві назвемо словники кінця ХІХ — початку ХХ ст. І. Войцеховича, Л. Боровиковського, А. Метлинського, П. Білецького-Носенка, О. Афанасьєва-Чужбинського, К. Шейковського, М. Закревського, Ф. Піскунова та ін., а також відомий “Словарь української мови” (1907-1909) Б.Грінченка.

³ *Мельничук О.С.* Словник української мови в 11-ти томах // Вісник АН УРСР. — 1984. — № 3. — С. 100-103; *Паламарчук Л.С.* Новий академічний словник // Мовознавство. — 1980. — № 5. — С. 3-9; *Паламарчук Л.С.* Тлумачний словник української мови в колі слов'янських словників цього типу. Доповідь на VII Міжнар. з'їзді славістів / Варшава, серпень 1973 р. — К.: Наук. думка, 1973. — 20 с.; *Ковалик І.І.* Зауваження до “Словника української мови” // Інформаційні матеріали наукової ради з проблеми “Закономірності розвитку національних мов у зв'язку з розвитком соціалістичних націй”. — К.: Наукова думка, 1974. — Вип. 16. — С. 50-53 та ін.

Вагомість для лінгвістичної науки 11-томного Словника української мови визначається трьома чинниками.

По-перше, він є підсумком розвитку вітчизняної мовознавства, в якому сконцентровано здобутки лінгвістичної теорії та практики декількох поколінь українських учених.

По-друге, він представляє фактичну базу для проведення нових мовознавчих досліджень. Зокрема, серед основних напрямів його дослідження назвемо аналіз реєстрового складу словника з граматичного погляду, стилістичного розмежування, а також характеристики основних параметрів лексикографічної інтерпретації лексики з позиції сучасності, пошук шляхів удосконалення фіксації лексичного складу мови та методів і способів подання семантичних структур та відношень. Назвемо праці⁴, які у зв'язку із СУМом не втратили свого значення і для нашого дослідження. Пріоритетним на сьогодні є модернізація СУМа в рамках програми створення Національної словникової бази — укладання на його основі нового 20-томного варіанту словника, а також його електронних відповідників для інформаційних комп'ютерних систем, створення різного роду семантичних аналізаторів та інших комп'ютерних систем, що застосовуються в контурах інтелектуальних мовно-інформаційних систем⁵.

По-третє, нові видання тлумачних словників, що з'явилися за останнє десятиліття, методологічно та інформативно також базуються на СУМі, по суті, представляючи його "клони". Принаймні, нам не вдалося знайти у них жодних принципів лексикографічних інновацій, незважаючи на подекуди значне (і часто-густо невинновдане) розширення реєстру.

У СУМі як одномовному тлумачному словнику план вираження слова, тобто ліва частина словникової статті, містить відображення фонемної, акцентуаційної, морфологічної (частиномовна приналежність, тип словозміни, граматичні категорії) характеристики, що

⁴ Паламарчук Л.С. Про принципи добору лексичного інвентаря до загальномовного словника // Мовознавство. — 1973. — № 3. — С. 3-11; Паламарчук Л.С. Українська радянська лексикографія: питання історії, теорії і практики. — К.: Наук. думка, 1978. — 203 с.; Слово і фразеологізм у словнику: Зб. наук. праць / Відп. ред. Л.С. Паламарчук. — К.: Наук. думка, 1980. — 250 с. Клименко Н.Ф., Пешак М.М., Савченко І.Ф. Формалізовані основи семантичної класифікації лексики. — К.: Наук. думка, 1982. — 252 с. Украинский семантический словарь. Проспект (машинный формат базы данных и принципы его автоматического составления) / М.М. Пешак, Н.Ф. Клименко, Е.А. Карпиловская и др. — К., Наук. думка, 1990. — 264 с. та багато ін.)

⁵ Широков В.А. Феноменология лексикографических систем. — К.: Наукова Думка, 2004. — 326 с.

становлять параметри граматичної семантики, а також стилістичної інформації про реєстрове слово. При цьому "фонетичні варіанти реєстрових слів, що вживаються паралельно, подаються в одній словниковій статті за алфавітом (ЖИТТЄВИЙ, ЖИТТЬОВИЙ)"⁶. Морфологічні характеристики наводяться як вказівка на тип словозміни або шляхом прив'язування реєстрової одиниці до відповідної частини мови. Морфематичні явища, викликані словозміною, подаються разом із словозмінними ознаками. Граматичні характеристики (вказівка на значення відповідних категорій, синтаксичні особливості) подаються після словозмінних і, як правило, взаємодіють з іншою лексикографічною інформацією словникової статті. Стилiстичні властивості лексичного значення реєстрового слова подаються у формі встановлених ремарок.

Вказівка на ступінь і характер обсягу лексичного значення (плану змісту), стилістично-функціональні особливості його розшарування зосереджується у правій частині словникової статті, визначаючи якісну значущість змісту, автономність та зв'язаність лексичного значення реєстрового слова у системі мови.

Обсяг і якісна специфіка індивідуального лексичного значення слів, уведених до реєстру Словника, зумовлює розбиття правої частини словникових статей на рубрики і підрубрики. У словникових статтях однозначних реєстрових слів рубрикація відсутня, принципі ж рубрикації лексичних значень полісемічних слів "дають змогу формувати словникові статті найрізноманітнішої складності. Лексичне значення розбивається на таку кількість рубрик і підрубрик, яку вимагає опрацьований фактичний матеріал"⁷.

Формули тлумачення (словникові дефініції), що подаються у рубриках (основні значення) та підрубриках (їх відтінки) "забезпечують певний автоматизм як при віднесенні граматично оформлених слів з лексичною семантикою мови, так і при користуванні словниками навіть у філологічно невідготовлених людей"⁸.

Окрім формул тлумачення, у правій частині словникових статей СУМу "для підтвердження існування слова в мові, для наочнішого і повнішого розкриття його значень, синтаксичних зв'язків, вживання в певному словесному оточенні й з певним стилістичним за-

⁶ Словник української мови / І.К. Білодід (гол. ред) та ін. — Т. 1-11. — К.: Наук. думка, 1970-1980. — Т. 1. — С. VIII.

⁷ Кліменко Н.Ф., Пещак М.М., Савченко І.Ф. Формалізовані основи семантичної класифікації лексики. — К.: Наук. думка, 1982. — 252 с. — С. 12.

⁸ Там само.

барвленням”⁹ широко використовуються ілюстрації. Останні в словниковій статті разом з формулами тлумачення включаються у відношення часткової кореляції, яка найбільш наочно проявляється в їх текстотвірній структурі: “нерідко синтаксичні зв’язки реєстрового слова в тексті ілюстрацій начебто віддзеркалюються в синтаксичних особливостях відповідника, вираженого формулою тлумачення, і не тільки в синтаксичних зв’язках, а й у конкретній лексичній сполучуваності”¹⁰.

“Формули тлумачення, супроводжувані ілюстраціями, репрезентують у словниковій статті глибинну структуру тієї частини висловлювання, яка має безпосереднє відношення до змісту лексичного значення реєстрової одиниці. Ілюстрації, таким чином, призначені представляти особливості вербалізованої структури реєстрових одиниць, а формули тлумачення — категоризовану структуру їх ролі у висловлюванні”¹¹.

Лексикографічний опис мовних одиниць як компонентів цілісної системи мови здійснюється, як уже зазначалося, за допомогою ремарок. Для Словника української мови прийнято понад 80 формальних ознак, які дають змогу тому, хто користується Словником, “не лише побачити розмаїтість і багатогранність словникового складу української літературної мови, але й дібрати найвлучніше слово для конкретної ситуації”¹². Якщо до формальних лексикографічних ремарок слід додати ще особливості подачі морфемної та словотворчої структури реєстрових одиниць і способів тлумачення, то отримуємо близько 100 вихідних формальних ознак, за допомогою яких інтерпретується лексична та граматична семантика реєстрових одиниць СУМа.

Відзначимо ще один аспект СУМа, перше звернення до якого відбулося саме у наших працях¹³, а саме — структурний аспект.

⁹ Словник української мови / І.К. Білодід (гол. ред) та ін. — Т. 1-11. — К.: Наук. думка, 1970-1980. — Т. 1. — С. XI.

¹⁰ Пешак М.М., Клименко Н.Ф., Ярун Г.М., Карпіловська Є.А. Лексична семантика в системі “людина — машина”. — К.: Наук. думка, 1986. — 282 с. — С. 276.

¹¹ Там само — С. 278.

¹² Паламарчук Л.С. Українська радянська лексикографія: питання історії, теорії і практики. — К.: Наук. думка, 1978. — 203 с. — С. 131.

¹³ Широков В.А., Воронько М.П., Костышин А.М. Структура и информационная модель толкового Словаря украинского языка. Труды научной конференции // “Проблеми створення машинних фондів мов”. — Київ, 1991. — С. 74-76. Широков В.А., Пешак М.М. Структурна модель реєстрової частини Словника української мови. Збірник наук. праць // “Національна архівна інформаційна система “Архівна та рукописна Україніка і комп’ютеризація архівної справи”, вип. 1: “Інформатизація архівної справи в Україні: сучасний стан та перспективи”. — К., 1996. — С. 154-174. Широков В.А. Інформаційна теорія лексикографічних систем, — К.: Довіра, 1998, 330 с.

