УДК 658.012.011.56

К.М. Якименко

ПРОГРАМНО-ТЕХНОЛОГІЧНІ АСПЕКТИ СТВОРЕННЯ ЛЕКСИКОГРАФІЧНОЇ СИСТЕМИ "СЛОВНИК УКРАЇНСЬКОЇ МОВИ"

Розглядається комп'ютерна технологія укладання нового тлумачного Словника української мови. Описано структуру лексикографічної бази даних (ЛБД) словника, принципи її побудови та внутрішні зв'язки між окремими елементами. Запропоновано клієнтську програму для редагування ЛБД словника, розглянуто її функціональні можливості.

Вступ

В 1999 році вийшов Указ Президента України "Про розвиток національної словникової бази" та згодом було затверджено план-графік реалізації визначених Указом завдань. Серед них своєю масштабністю вирізняється створення нової версії багатотомного тлумачного Словника української мови. Його обсяг планується на рівні 20 томів, а прототипом слугує 11-томний академічний Словник української томи (СУМ) [1]. Робота над створенням СУМа тривала понад 30 років, і він становить цілу епоху у вітчизняній лінгвістиці, подаючи в цілісному вигляді всю систему української мови, яка склалася.

Було усвідомлено, що лексикографічний проект такого масштабу й рівня не може здійснюватися без застосування сучасних комп'ютерних технологій на всіх етапах лексикографічного опрацювання мовних одиниць. Іншим вирішальним аргументом щодо використання цих технологій виявилася необхідність створення сучасної національної словникової бази у надзвичайно стислі терміни (4 роки), що нереально за використання традиційних методів лексикографування.

Отже, проект нового тлумачного Словника української мови передбачує створення щонайменше двох кінцевих продуктів: традиційного паперового багатотомного словника, "відправною точкою" якого слугує 11-томний СУМ, і комп'ютерної лексикографічної системи, що не лише вбирає зміст паперового варіанта словника, але й містить цілу низку додаткових інформаційних і лінгвістичних функцій. Така постановка завдання є новою для вітчизняного словникарства, тому вимагає вироблення концептуальних засад, які б враховували разом із суто лінгвістичними аспектам також організаційні та технологічні (а отже, й програмні) проблеми реалізації фундаментального лексикографічного процесу.

Технологія створення лексикографічної бази даних СУМа

Розробка структури фундаментальної академічної лексикографічної системи "Словник української мови", створення лексикографічних баз даних (ЛБД) тлумачного типу та клієнтського програмного забезпечення опрацювання цих ЛБД розпочалося з обробки тексту СУМа на основі його структурної теорії [2]. З метою створення ЛБД СУМа було здійснено конверсію паперового варіанту 11-томника [1] до електронної форми. Цей етап виконувався засобами сканування та розпізнавання тексту, в результаті чого було одержано цифровий варіант тексту 11-томного СУМа. Ланцюг підготовки тексту СУМа до конвертації в ЛБД представлено на рис. 1. Після сканування і розпізнавання тексту СУМа (9856 сторінок) його було збережено в RTF-форматі та роздруковано для коректури з метою виправлення помилок, які виникли при роботі програми оптичного розпізнавання. Після подвійної коректури й внесення виправлень до електронного тексту СУМа було одер-

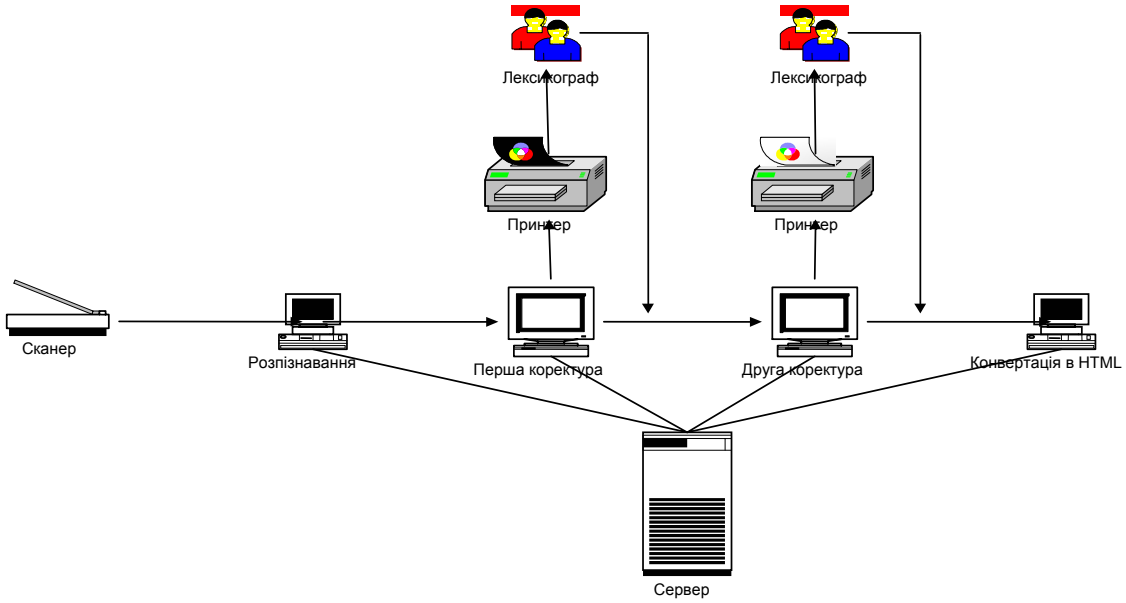

Рис. 1. Ланцюг підготовки тексту СУМа до конвертації в ЛБД

жано еталонний варіант тексту 11-томника. Останній етап підготовки до створення ЛБД — конверсія з RTF- до HTML-формату з системою кодування Unicode засобами текстового редактора MS Word.

Слід відзначити, що робота з ЛБД у форматі електронного тексту такого значного обсягу (понад 135 Мбайт дискового простору) є абсолютно неефективною. Окрім дуже повільного темпу, з яким система працює на таких текстових масивах, принципову незручність представляє неможливість прямого доступу до окремих елементів структури словника. Такі вади годні звести нанівець усі переваги, які надає використання комп'ютерів у лексикографії. Отже необхідною постала розробка спеціалізованої ЛБД Словника української мови, причому її формування треба було здійснити автоматично, оскільки в ручному або напівручному режимі створення такої бази даних, як переконує досвід, є просто неможливим.

Для забезпечення процесу автоматичної конверсії тексту СУМа до ЛБД в Українському мовно-інформаційному фонді НАН України було розроблено програмне забезпечення виділення елементів його структури відповідно до будови Л-системи та з використанням поліграфічних ознак її текстової ідентифікації. В результаті конверсії весь текст СУМа з HTML-файлів був перенесений до ЛБД зі структурою, яку розглянемо детальніше.

Структура лексикографічної бази даних СУМа

Структура ЛБД СУМа є мінімальною — вона дозволяє правильно відображати всі структурні елементи СУМа, але водночас має можливості для розширення. Наприклад, ліва частина словникової статті у цій ЛБД не є структурованою, а записується єдиним блоком.

Структура ЛБД СУМа разом з клієнтською програмою дозволяє не тільки представити, а й візуалізувати представлення будь-якої словникової статті СУМа у вигляді дерева. При цьому значно спрощується доступ до структурних елементів статті, а зв'язки між елементами унаочнюються. Чимало допоміжних елементів статті (термінальні символи, номери тлумачень, певні розділові знаки, спеціальні символи та шрифтові виділення) не потребують збереження у ЛБД, а можуть додаватися динамічно під час формування статті програмою виводу. Такі автоматичні операції допомагають уникнути багатьох помилок при редагуванні статей, а можливість помилкового введення еле-

ментів, які порушують структуру словника, взагалі виключається. Процес редагування стає більш простим, контрольованим та уніфікованим, легко здійснюються операції додавання, вилучення та виправлення елементів словникових статей. Гнучкість структури ЛБД СУМа дозволяє в перспективі реалізувати те, що було принципово неможливим при представленні СУМа у вигляді послідовного тексту. Наприклад, якщо фрагменти різних статей повторюються, то такий фрагмент можна зберігати в базі тільки один раз, але при цьому створити покажчики, які пов'язували б цей фрагмент з усіма відповідними статтями. Такий механізм разом з уже реалізованими функціями динамічного формування приводить до економії дискового простору.

У структурі ЛБД СУМа виділено сукупність таблиць (див. табл. 1 — 10) та зв'язків між ними, зміст яких розкривається нижче.

- **ID** — унікальний ідентифікатор статті;
- **Reestr** — реєстрове слово СУМа;
- **Part** — код частини мови реєстрового слова;
- **Data** — дата та час, коли робилася остання правка статті;
- **Digit** — цифровий код реєстрового слова, що використовується для сортування (в цьому коді літери українського алфавіту позначені двома цифрами: А — 01, Б — 02, В — 03 і т. д., числа позначаються чотирма цифрами: 1 — 0001, 2 — 0002, ..., 10 — 0010 і т. д. Всі інші символи ігноруються);
- **IsLink** — ознака, чи є стаття відсилковою (при виводі позначається "*див.*");
- **LinkText** — текст відсилки;
- **IsOldSum** — ознака, чи належить стаття до 11-томного СУМу або є новою;
- **IsDel** — ознака, чи є стаття видаленою з БД;
- **QtyEd** — кількість правок статті;
- **FinalEd** — ознака, чи була стаття узгоджена з головним науковим редактором;
- **NREd** — ознака, чи була стаття узгоджена з науковим редактором;
- **Printed** — ознака, чи була стаття відправлена на роздрук;
- **Odious** — ознака, чи належить стаття до проблемних (таких, що мають недоліки в інтерпретаційній частині та підлягають подальшому перегляду та редагуванню).

Таблиця проіндексована за полями: **ID** (Unique), **Reestr**, **Part**, **Digit**.

Таблиця 1. Реєстрові слова СУМа ("**nom**")

ID	Reestr	Part	Data	Digit	IsLink	LinkText	IsOldSum	IsDel	QtyEd	FinalEd	NREd	Printed	Odious
31	АБИ#ЯКИЙ	3	18.03.02 15:35	1021133151114	0		1	0	1	1	1	1	0
32	АБІОГЕНЕ#З	1	18.03.02 15:35	10212190407180700	0		1	0	1	1	1	1	0
33	АБІОГЕ#ННИЙ	3	18.03.02 15:36	1021219040718180000	0		1	0	1	1	1	1	0
34	АБІСА#ЛЬ	1	18.03.02 15:37	1021222011631	0		1	0	1	1	1	1	0
35	АБІСА#ЛЬНИЙ	3	19.03.02 10:11	1021222011631180000	0		1	0	2	1	1	1	0
36	АБІССІ#НЕЦЬ	1	19.03.02 10:12	1021222221218070000	1	абіссі#нці.	1	0	2	1	1	1	0
37	АБІССІНКА	1	19.03.02 10:12	10212222212181500	1	абіссі#нці.	1	0	3	1	1	1	0
38	АБІССІ#НСЬКИЙ	3	19.03.02 10:12	10212222212182200000000	0		1	0	2	1	1	1	0
39	АБІССІ#НЦІ	21	19.03.02 10:12	10212222212182700	0		1	0	1	1	1	1	0
40	АБІТУРІЄ#НТ	1	19.03.02 10:14	1021223242112080000	0		1	0	1	1	1	1	0

Таблиця 2. Діапазони редагування СУМа ("**Ranges**")

Part	Lower	Upper	Letter	LexEd
1.1			Й	Горюшина Г.Н.
1.1	Метро	мозаїка		Горюшина Г.Н.
1.2	М	метрівка		Лозова Н.Є.
1.2	Р	розвиватися		Лозова Н.Є.
1.3			Д	Самойлова І.А.
1.3	мозковий	м'ячик		Самойлова І.А.
2.1	П	переступник		Бибик С.П.
2.2	літо	льяноси		Єрмоленко С.Я.
2.2	рух	ряхтливий		Єрмоленко С.Я.
2.3	розвиднитися	рутяний		Неровня Н.М.

- **Part** — номер технологічного тому (Т-тому) чи його частини;
- **Lower** — слово, яке є нижньою границею частини;
- **Upper** — слово, яке є верхньою границею частини;
- **Letter** — літера, яка цілком належить до частини тому (якщо це поле непусте, то значення полів **Lower** та **Upper** ігноруються та навпаки);
- **LexEd** — ім'я та ініціали наукового редактора або лексикографа, відповідального за Т-том або частину Т-тому.

Таблиця 3. Ліві частини статей ("**lr**")

ID	Left	Right	IsDel
1			0
2			1
3			0
4			0
5			0

- **ID** — унікальний ідентифікатор статті (має збігатися з відповідним ідентифікатором статті з таблиці **nom**);
- **Left** — текст лівої частини;
- **Right** — зарезервовано для правої частини статті; не використовується;
- **IsDel** — ознака, чи є запис видаленим з БД;

Таблиця проіндексована за полем **ID** (Unique).

Таблиця 4. Блоки тлумачень ("**intgroup**")

ID	ID_lv	NumbGr	IsDel	Param
1	168538	1	1	Протиставний
2	168538	2	1	Зіставний
3	168539	1	0	протиставний.
4	168539	2	0	зіставний.
5	168539	3	0	приєднувальний.
6	168539	4	0	приєднально-підсилювальний; у сполуч. із займ. і присл.</I> як, який, скільки, що <I>та ін.
7	168539	5	0	єднальний, діал.

- **ID** — унікальний ідентифікатор блока;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язаний блок;
- **NumbGr** — номер блока в межах статті (нумерація має бути послідовною);
- **IsDel** — ознака, чи є запис видаленим з БД;
- **Param** — параметр блока.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbGr**.

Таблиця 5. Тлумачення ("**interpr**")

ID	ID_lv	Relat	NumbInt	IsDel	Lv
7459	6305	1	18	0	0
7460	6305	1	19	0	0
7461	6305	11	20	0	0
7462	6305	11	21	0	0
7463	6305	11	22	0	0
7464	6305	9	23	0	0
7465	6305	9	24	0	0
7466	6305	9	25	0	0
7467	6305	9	26	0	0
7468	6305	5	27	0	0

- **ID** — унікальний ідентифікатор тлумачення;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язане тлумачення;
- **Relat** — код типу відношення, яке виражають сполучення з реєстровим словом у даному тлумаченні (тіль-

ки для прийменникових словникових статей);

- **NumbInt** — номер тлумачення в межах фрагмента (нумерація має бути послідовною);
- **IsDel** — ознака, чи є запис видаленим з БД;
- **Lv** — код типу фрагмента вищого рівня.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbInt, Lv**.

Таблиця 6. Фразеологізми та еквіваленти слів ("**fraseol**")

ID	ID_lv	NumbFras	Kind	Fras	IsDel	Lv
6	65	1	5		0	2
7	79	1	2		0	2
8	83	1	2		0	2
9	84	1	2		0	2
10	119	1	5		0	2
11	119	2	4		0	2
12	119	3	4		0	2
13	119	4	4		0	2
14	119	5	4		0	2
15	119	6	4		0	2

- **ID** — унікальний ідентифікатор фразеологізму або еквівалента слова;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язаний фразеологізм або еквівалент слова;
- **NumbFras** — номер фразеологізму або еквівалента слова в межах фрагмента (нумерація має бути послідовною);
- **Kind** — вид фразеологізму або еквівалента слова;
- **Fras** — назва фразеологізму або еквівалента слова;
- **IsDel** — ознака, чи є запис видаленим з БД;
- **Lv** — код типу фрагмента вищого рівня.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbFras, Lv**.

Таблиця 7. Відтінки ("**shade**")

ID	ID_lv	NumbShade	Lv	IsDel
5	7	2	2	1
6	7	3	2	1
7	7	4	2	1
8	7	5	2	1
9	7	6	2	1
10	8	1	2	1
11	42	1	2	0
12	49	1	2	0
13	49	2	2	0
14	49	3	2	0

- **ID** — унікальний ідентифікатор відтінка;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язаний відтінок;
- **NumbShade** — номер відтінка в межах фрагмента (нумерація має бути послідовною);
- **Lv** — код типу фрагмента вищого рівня;
- **IsDel** — ознака, чи є запис видаленим з БД.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbShade, Lv**.

Таблиця 8. Частини тлумачень або значення фразеологізмів (еквівалентів слів) ("**subshade**")

ID	ID_lv	NumbSub	Lv	IsDel
2726	171923	1	2	0
2727	171923	2	2	0
2728	41053	1	4	0
2729	41053	2	4	0
2730	16581	1	3	0
2731	16581	2	3	0
2732	102061	1	2	0
2733	102061	2	2	0
2735	1587	2	3	0
2736	32095	1	3	0

- **ID** — унікальний ідентифікатор частини тлумачення або значення;

- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язана частина тлумачення або значення;
- **NumbSub** — номер частини тлумачення або значення в межах фрагмента (нумерація має бути послідовною);
- **Lv** — код типу фрагмента вищого рівня;
- **IsDel** — ознака, чи є запис видаленим з БД.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbSub, Lv**.

Таблиця 9. Формули тлумачень ("**formula**")

ID	ID_lv	NumForm	Interpr	Lv	Paradigm	Vid	Perexidn	Keruvan	Spoluch	Rid	Chislo	Style	ElsOll	IsDel
114	8	1		3										0
115	84	1		2									мед.	0
116	7	1		5										0
117	8	1		5										0
118	85	1		2									біол.	0
119	86	1		2									кого, мед.	0
120	87	1		2									біол.	0
121	88	1		2										0

- **ID** — унікальний ідентифікатор формули тлумачення;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язана формула тлумачення;
- **NumForm** — номер формули тлумачення в межах фрагмента;
- **Interpr** — текст формули тлумачення;
- **Lv** — код типу фрагмента вищого рівня;
- **Paradigm** — парадигматичний клас;
- **Vid** — вид;
- **Perexidn** — перехідність;
- **Keruvan** — керування;
- **Spoluch** — сполучуваність;
- **Rid** — рід;
- **Chislo** — число;
- **Style** — стиль;
- **ElsOll** — інші параметри формули тлумачення;
- **IsDel** — ознака, чи є запис видаленим з БД.

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumForm, Lv**.

Таблиця 10. Ілюстрації ("**illustr**")

ID	NumbIll	Illustr	Author	Title	Edition	Pages	Figur	Cm	IsDel	ID_lv
595	3		А. Малишко		0	0	0	0	0	558
596	1		А. Шиян		0	0	0	0	0	560
597	2		Леся Українка		0	0	0	1	0	560
598	1		з газ.		0	0	0	0	0	561
599	2		В. Собко		0	0	0	0	0	561
600	3		Остап Вишня		0	0	1	0	0	561
601	1				0	0	0	0	0	562
602	1				0	0	0	0	1	563
603	1		Український світ		0	0	0	0	1	564
604	1		І. Франко		0	0	0	0	0	565

- **ID** — унікальний ідентифікатор ілюстрації;
- **NumbIll** — номер ілюстрації в межах формули тлумачення (нумерація має бути послідовною);
- **Illustr** — текст ілюстрації;
- **Author** — автор джерела ілюстрації;
- **Title** — назва джерела ілюстрації (введено для сумісності з 11-томним СУМом, у новому СУМі не використовується);
- **Edition** — рік видання джерела ілюстрації (введено для сумісності з 11-томним СУМом, у новому СУМі не використовується);
- **Pages** — номер сторінки з джерела ілюстрації (введено для сумісності з 11-томним СУМом, у новому СУМі не використовується);
- **Figur** — ознака, чи реєстрове слово вживається в ілюстрації в образному значенні;
- **Cm** — ознака, чи реєстрове слово вживається в ілюстрації у порівнянні;

- **IsDel** — ознака, чи є запис видаленим з БД;
- **ID_lv** — ідентифікатор фрагмента вищого рівня, з яким пов'язана ілюстрація;

Таблиця проіндексована за полями: **ID** (Unique), **ID_lv**, **NumbIll**.

Перейдемо до опису зв'язків між таблицями ЛБД СУМ, представлених на рис. 2. Головною таблицею ЛБД є таблиця **nom**, кожний її запис відповідає одній словниковій статті СУМа. Таблиця **lr** зв'язана з nom за принципом "один до одного", тому що кожна стаття має тільки одну ліву частину. Зв'язок здійснюється через поле **ID**. Таблиці **intgroup, interpr, fraseol, shade, subshade, formula** та **illustr** мають у своєму складі поля **ID**, **ID_lv** та **Lv** (крім intgroup та illustr, які не мають поля Lv). Ці поля використовуються для зв'язків між таблицями, і їх призначення у всіх перерахованих таблицях однакове. Поле **ID** є унікальним ідентифікатором запису; **ID_lv** — ідентифікатор вищого рівня, з яким пов'язаний даний запис; **Lv** визначає, яка саме таблиця (тобто тип фрагмента статті) вищого рівня мається на увазі. Наприклад, якщо в деякому запису в таблиці interpr Lv має значення 0, то цей запис (тлумачення) відноситься до реєстрового слова (а не є частиною блока тлумачень), такого, що його ID дорівнює значенню ID_lv для цього запису.

Далі для кожної таблиці перераховані ті, які можуть бути для неї безпосереднім вищим рівнем:

- **intgroup**: nom;
- **interpr**: nom, intgroup;
- **fraseol**: nom, interpr;
- **shade**: interpr, fraseol, subshade (тільки коли він відноситься до fraseol);
- **subshade**: interpr, fraseol, shade (тільки коли він відноситься до interpr);
- **formula**: interpr, fraseol, shade, subshade;
- **illustr**: formula.

Як видно з цього переліку, поля Lv не мають ті таблиці, які зв'язані тільки з однією таблицею.

Значення, які приймає Lv залежно від таблиці вищого рівня:

0	nom	3	fraseol
1	intgroup	4	shade
2	interpr	5	subshade

Зв'язки між таблицями ЛБД СУМ відповідно до структури СУМа можна інтерпретувати наступним чином:

⇒ Словникова стаття СУМа завжди має ліву частину.

⇒ Стаття може складатися з декількох (1 та більше) блоків тлумачень, тлумачень, фразеологізмів або еквівалентів слова (у випадку, якщо слово вживається тільки у складі фразеологізму або еквівалента слова).

⇒ Блок може складатися з декількох тлумачень.

⇒ Тлумачення може мати відтінки, частини тлумачення та пов'язані з ним фразеологізми або еквіваленти слова.

⇒ Фразеологізм та еквівалент слова може мати декілька значень та відтінки.

⇒ Значення фразеологізму або еквівалента слова також може мати відтінки.

⇒ Тлумачення, відтінок, частина тлумачення, фразеологізм (еквівалент слова), значення фразеологізму (еквівалента слова) мають у своєму складі формулу тлумачення (інколи вони можуть її не мати, але тоді вважається, що є фіктивна (пуста) формула тлумачення) та можуть мати ілюстрації.

⇒ Формула тлумачення фактично є компонентом тлумачення, відтінку, частини тлумачення, фразеологізму (еквівалента слова) або значення фразеологізму (еквівалента слова). Але структури всіх цих формул тлумачення подібні, тому вони винесені в окрему таблицю **formula**. Як видно зі структури цієї таблиці (див. табл. 9), вона зберігає як безпосередньо текст формули тлумачення, так і її граматичні, стилістичні та інші параметри. Для цих параметрів зарезервовано окремі поля, але на сьогодні усі параметри зберігаються в полі **ElsOll**.

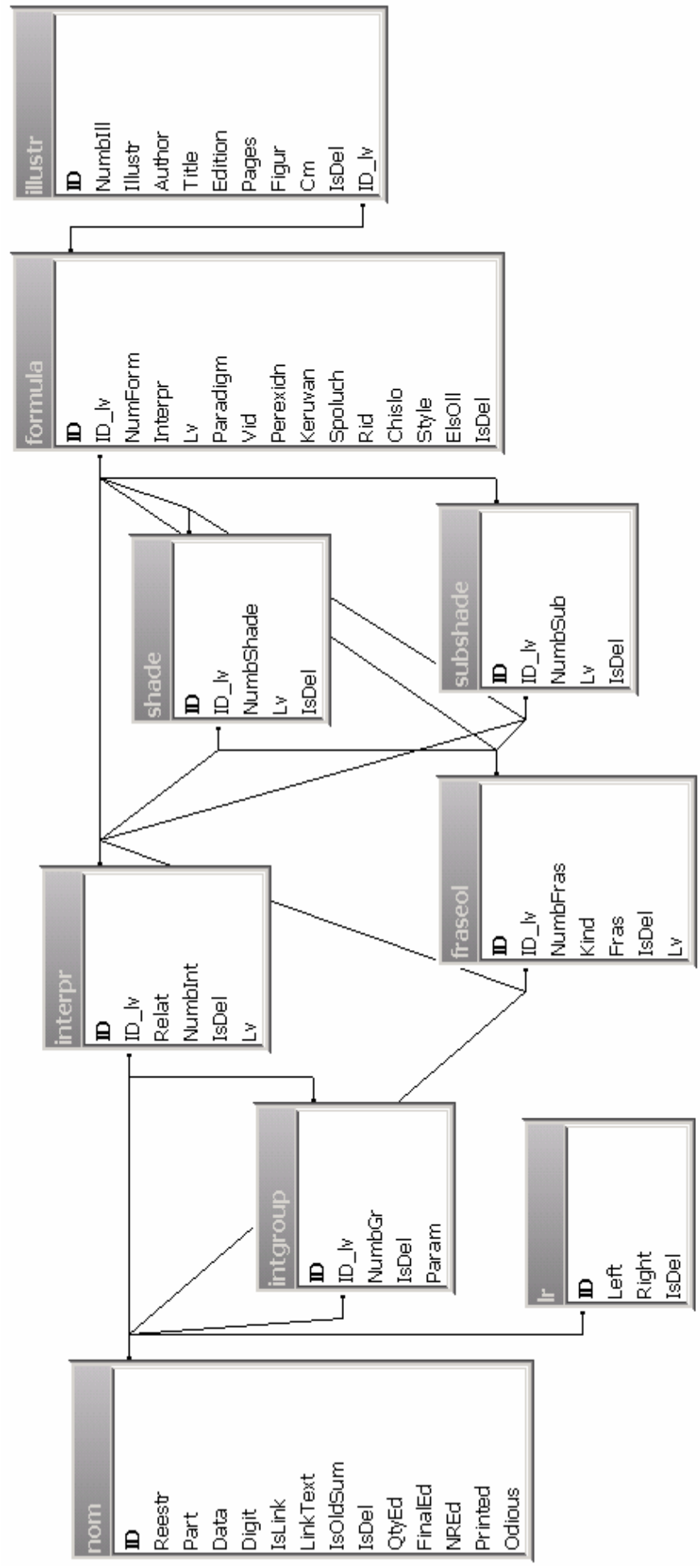

Рис. 2. Схема зв'язків між таблицями ЛБД СУМ

⇒ Текстові ілюстрації з прикладами вживання реєстрового слова у певних значеннях можуть бути наявні тільки тоді, коли є формула тлумачення для відповідного значення. Тому таблиця ілюстрацій **illustr** зв'язана тільки з таблицею формул тлумачення formula. Можливі випадки, коли формули тлумачення фактично немає, а є тільки параметри — тоді в таблиці formula створюється запис з нульовим значенням поля **Interpr**. Фразеологізм та еквівалент слова також можуть не мати формули тлумачення — у випадках, коли вони є відсилковими або мають декілька значень. В обох цих випадках відповідний йому запис у таблиці formula може не створюватись, тому такий фразеологізм чи еквівалент слова не має власних ілюстрацій.

Структури таблиць **intgroup**, **interpr**, **fraseol**, **shade** та **subshade** дуже подібні між собою, але мають деякі відмінності:

- **intgroup**: блок не має формули тлумачення, але може мати параметри. Тому його параметри є безпосередньою частиною таблиці intgroup та записуються в поле **Param;**
- **interpr**: тлумачення може характеризуватися відношенням, код якого записується у поле **Relat**. Для тлумачень з однаковим типом відношення, які належать до одного вищого рівня та мають послідовні номери, при формуванні статті тип вказується тільки один раз для першого з групи тлумачень;
- **fraseol**: фразеологізм завжди має назву (тобто текст самого фразеологізму, до якого обов'язково входить реєстрове слово), яка записується у поле **Fras**. Також фразеологізм має поле **Kind** для виду фразеологізму, що позначається числом від 1 до 5. Вид визначає, як фразеологізм буде позначений при виводі статті. Фразеологізми однакового виду, що відносяться до одного вищого рівня та мають послідовні номери, поєднуються в блок, позначка до якого виводиться тільки один раз. Фразеологізм виду 5 не має ніякої спеціальної позначки. Все те саме справедливе також для еквівалентів слів, тільки поле **Kind** для них набуває значення 6;
- **subshade**: ця таблиця не має відзнак у структурі, але її особливість полягає в тому, що вона зберігає як частини тлумачень та відтінків, так і значення фразеологізмів. При цьому вважається, що коли фразеологізм (або його значення) має відтінок, то такий відтінок не може мати частин, інакше це призводило б до виникнення циклу в структурі зв'язків таблиць.
- Таблиця **illustr** має поле **Illustr** з безпосередньо текстом ілюстрації, а також поле **Author** для збереження прізвища автора або назви джерела ілюстрації. Поля **Title**, **Edition** та **Pages** створені виключно для сумісності з 11-томним СУМом і в новому СУМі не використовуються, але якщо вони заповнені, то при формуванні статті виводяться саме в такій послідовності (про поля **Figur** та **Cm** див. опис таблиці **illustr)**.

Таблиці **intgroup**, **interpr**, **fraseol**, **shade**, **subshade**, **formula** та **illustr** мають поля для внутрішньої послідовної нумерації записів у рамках статті або її фрагменту. Ці поля відповідно такі для таблиць:

- intgroup: **NumbGr**;
- interpr: **NumbInt**;
- fraseol: **NumbFras**;
- shade: **NumbShade**;
- subshade: **NumbSub**;
- formula: **NumForm**;
- illustr: **NumbIll**.

Формула тлумачення фактично завжди тільки одна, тому значення поля **NumForm** завжди має дорівнювати 1. Це поле є надлишковим, але його введено до таблиці для більшої одноманітності структур.

При додаванні нових записів до таблиць мають виконуватися наступні умови.

Значення **ID** для нового запису має дорівнювати найбільшому з існуючих значень плюс одиниця (це правило дійсне для всіх таблиць, крім **lr**). Значення **IsDel** має дорівнювати 0. Інші

умови розглянемо для кожної таблиці окремо.

⇒ Таблиця **nom**. Значення **Digit** визначається відповідно до реєстрового слова (див. опис таблиці **nom**). **Part** заповнюється згідно з таблицею кодів частин мови, яка в даній статті не приводиться. У поле **Data** заноситься поточний час. Поля-ознаки мають бути встановлені для нового запису наступним чином:

⇒ **IsOldSum** = 0;
⇒ **QtyEd** = 0;
⇒ **FinalEd** = 0;
⇒ **NREd** = 0;
⇒ **Printed** = 0;
⇒ **Odious** = 0.

⇒ Таблиця **lr**. Значення **ID** має дорівнювати ідентифікатору реєстрового слова ID, з яким пов'язана ліва частина.

⇒ Таблиці **intgroup, interpr, fraseol, shade, subshade, formula, illustr**. **ID_lv** має дорівнювати **ID** запису вищого рівня, з яким пов'язаний запис, що додається. **Lv** заповнюється так, як було описано вище (крім **illustr**). Для визначення значення поля внутрішньої нумерації (крім **formula**) вибираються всі поля з цієї таблиці, які мають ті самі значення ID_lv та Lv, тобто пов'язані з тим самим записом вищого рівня. З них вибирається найбільше значення поля **Num**XXX та до нього додається одиниця.

⇒ Таблиця **interpr**. Значення типу відношення **Relat** за умовчуванням дорівнює 0.

⇒ Таблиця **fraseol**. Якщо додається фразеологізм, його значення виду **Kind** за умовчуванням дорівнює 5, якщо еквівалент слова — то 6.

⇒ Таблиця **formula**. Звичайно новий запис для неї створюється одночасно з додаванням запису до таблиці **interpr**, **fraseol**, **shade** або **subshade**, з якою пов'язана формула тлумачення, і визначаються відповідні ID_lv та Lv. Значення поля **NumForm** має дорівнювати 1.

Фрагмент статті видаляється з БД без збереження будь-яких даних про нього, а якщо вилучається ціла стаття, то її фрагменти та саме реєстрове слово насправді не видаляються з БД, а значення ознаки IsDel для них встановлюється рівним 1.

При вилученні будь-якого фрагменту мають бути вилучені також усі пов'язані з ним елементи нижчих рівнів. Наприклад, на рівні видалення формули тлумачення відбувається також пошук та видалення всіх пов'язаних з нею ілюстрацій.

Програмний комплекс редагування ЛБД СУМа

ЛБД СУМа, реалізована в УМІФі, функціонує під СКБД Microsoft SQL Server 7.0. Клієнтську програму редагування ЛБД СУМа було розроблено і створено в середовищі Microsoft Visual Studio 6.0, вона працює під операційною системою Microsoft Windows 2000 або Microsoft Windows XP.

Програма орієнтована на роботу в мережевому середовищі, де багато користувачів одночасно мають доступ до ЛБД СУМа. У цьому випадку залежно від пріоритету користувачі можуть отримати доступ до всієї бази або її частини, можливість редагування статей або тільки їх перегляду. Крім того, оскільки для редагування реєстр СУМа був розбитий на 9 приблизно рівних технологічних томів, за кожний з яких відповідає окремий науковий редактор, а кожний з цих томів у свою чергу поділений між 3 або 4 лексикографами, було вирішено ввести діапазони редагування СУМа безпосередньо в ЛБД до таблиці Ranges (див. табл. 2). Для кожного з 9 томів існує свій логін, що обмежує доступ технолога, який працює з БД через клієнтську програму, тільки даним конкретним томом і блокує доступ до інших. Для контролю за процесом редагування в ЛБД введено кілька спеціальних полів (див. табл. 1). В режимі повного доступу також може бути здійснена фільтрація реєстру за технологічним томом або його частиною, за яку відповідає окремий лексикограф.

Клієнтська програма реалізує багато функцій для роботи з ЛБД. Ці функції виконуються окремими модулями

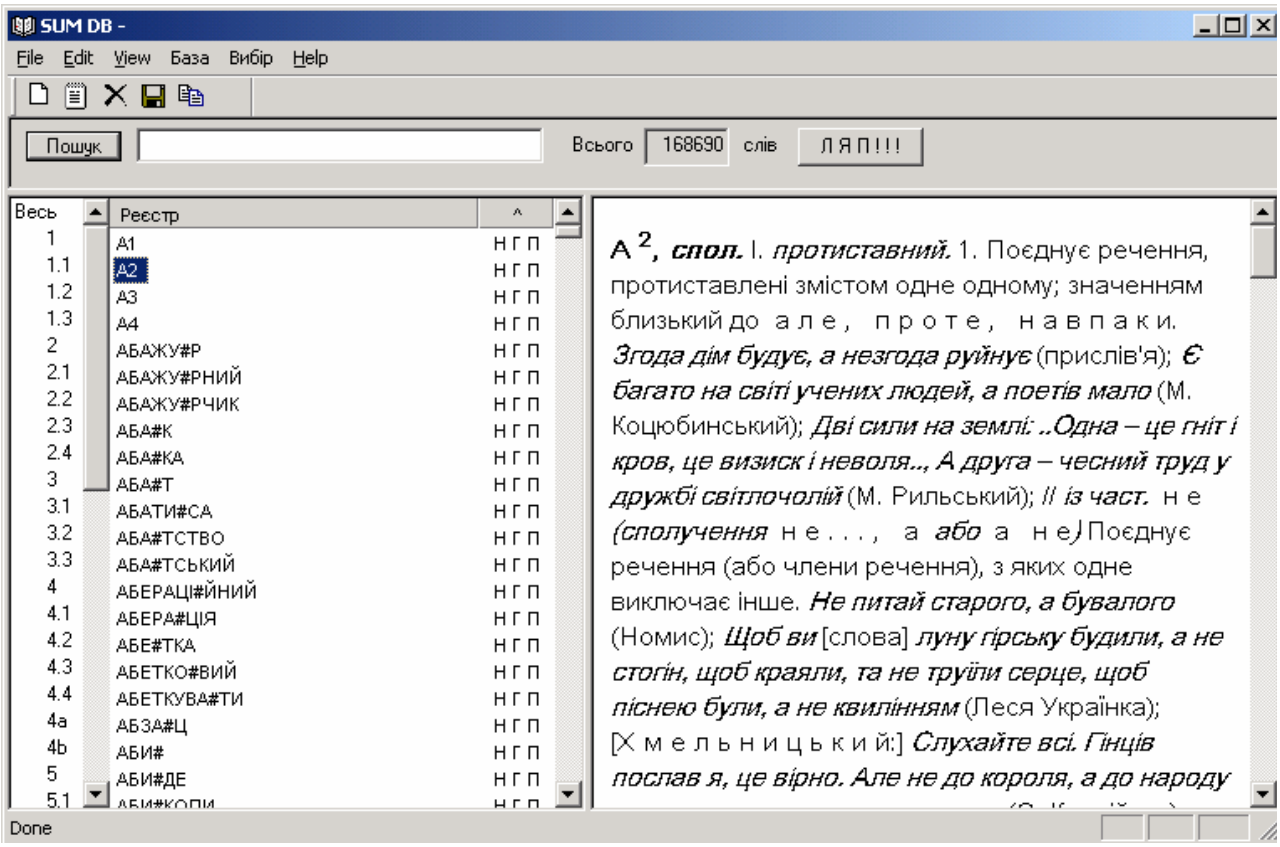

Рис. 3. Головне вікно програми

програми. Далі наведено стислий опис модулів та приклади зовнішнього вигляду діалогових вікон, пов'язаних з ними (рис. 3 – 8).

DicUASplApp — головний модуль програми. Містить функції для роботи зі статтею в цілому: додавання, вилучення, копіювання, перехід до редагування статті, встановлення ознак редагування, запис статей у файл для наступного роздруку, а також функції, пов'язані з переглядом: встановлення шрифту, вибір режиму фільтрації (за частиною мови, діапазонами редагування, довільним запитом). Крім того, до модуля входить декілька системних функцій: перевірка належності слова до потрібного діапазону, перетворення слова в його цифровий код, реакція на натиснення певних клавіш та інші; також тут описані основні глобальні змінні програми.

DicUASplView, **TrGr**, **FindForm** — головне вікно програми. Сюди входять функції вибору діапазону редагування та статті для перегляду, пошуку реєстрового слова, а також системні функції ініціалізації бази даних СУМа та дерева діапазонів редагування.

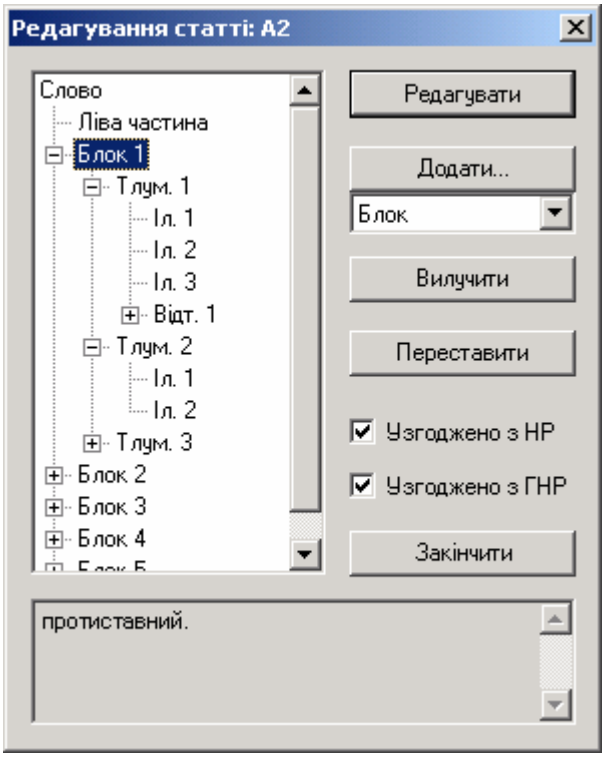

Рис. 4. Перегляд структури статті

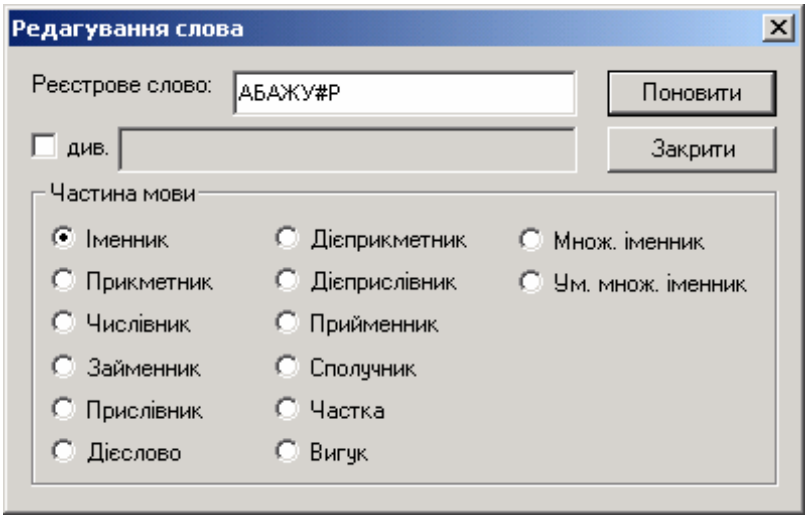

Рис. 5. Редагування слова

ArtTree — перегляд структури статті. Звідси можуть бути викликані функції додавання та вилучення окремих елементів статті, редагування їх, переставлення у потрібній послідовності та встановлення ознак редагування. Якщо програму викликано в режимі тільки перегляду, то можна здійснювати навігацію по структурі статті, але функції редагування при цьому будуть недоступні.

WordE — редагування слова. Дає можливість ввести реєстрове слово та указати, до якої частини мови воно належить. Якщо стаття є відсилковою, тут встановлюється відповідна ознака та вводиться реєстрове слово, на яке вона посилається.

InterprE — редагування тлумачення. У цьому вікні вводиться значення формули тлумачення та, при необхідності, встановлюється потрібне відношення. Може бути здійснена навігація між тлумаченнями, що належать до однієї словникової статті або блока. Аналогічні діалогові вікна (з деякими відмінностями) є також для редагування блоків тлумачень (**GroupE**), фразеологізмів (**FraseolE**), відтінків (**ShadeE**), частин тлумачень та значень фразеологізмів (**SubE**), ілюстрацій (**IllustrE**).

AddParam — редагування додаткових параметрів. У цьому вікні можна ввести значення граматичних, стилістичних та інших параметрів формули тлумачення.

WriteHTML — запис статей у файл. Статті можуть бути вибрані з діапазону (задається початкове та кінцеве слово) або зі списку. Якщо

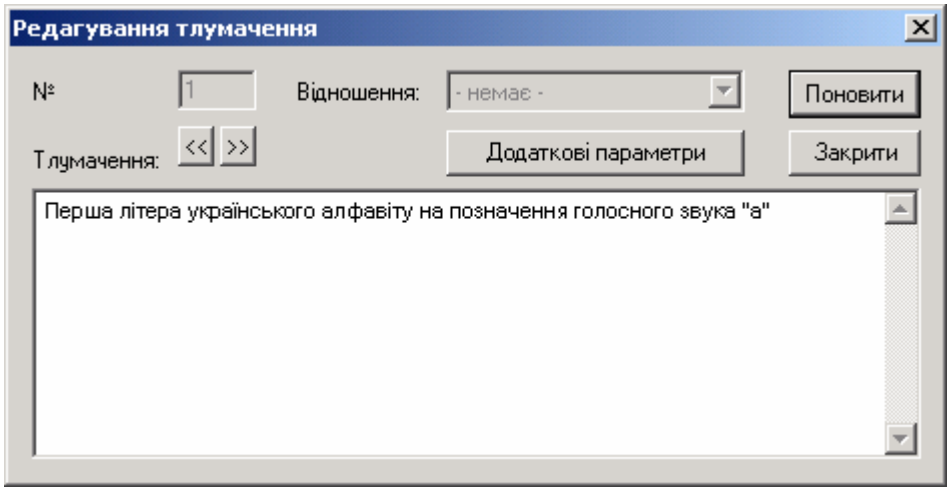

Рис. 6. Редагування тлумачення

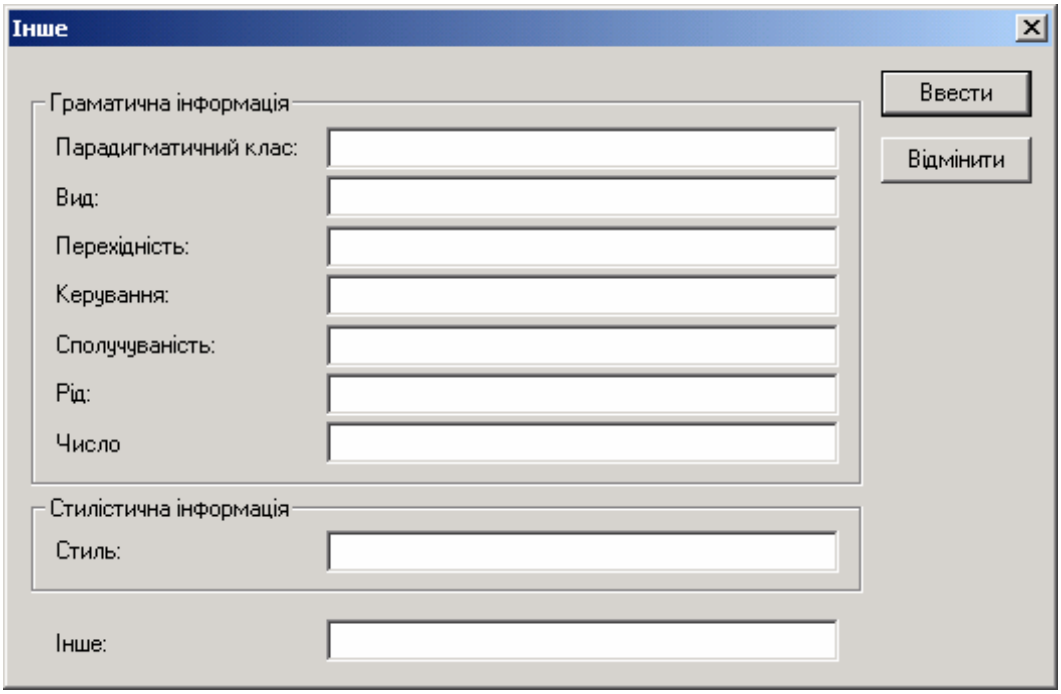

Рис. 7. Редагування додаткових параметрів

користувач має доступ тільки до певного діапазону редагування, то він може вибрати статті тільки в межах цього діапазону. Також вибирається ім'я файла (формату HTML), до якого будуть записані статті.

MakeHTML — модуль функцій для формування словникової статті в HTML-форматі, зокрема її елементів у тому вигляді, у якому вони мають бути представлені на екрані. Безпосередньо MakeHTML викликає ці функції в циклі та формує зовнішній вигляд статті згідно з усіма потрібними шрифтовими виділеннями та іншими поліграфічними ознаками.

Висновок

Розроблена в УМІФі лексикографічна база даних Словника української мови та програмний комплекс її редагування дозволяють організувати процес створення нового Словника значно ефективніше, ніж це було б

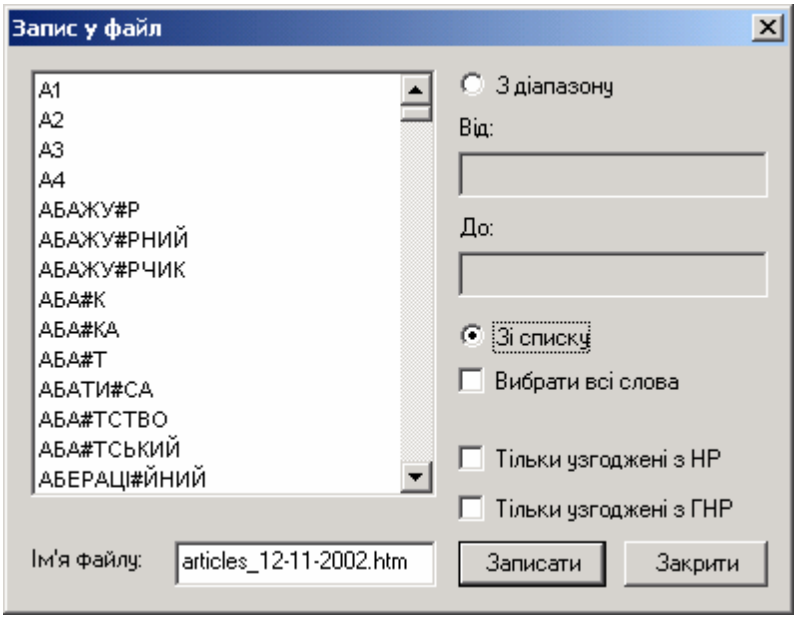

Рис. 8. Запис статей у файл

можливо при використанні тільки традиційних засобів укладання словників, та максимально підвищити продуктивність лексикографічної праці.

1. *Словник* української мови в 11-и томах. — К.: Наук. думка, 1970—1980.
2. *Широков В.А.* Інформаційна теорія лексикографічних систем. — К.: Довіра, 1998. — 331 с.

Отримано 06.11.03

Про автора

Якименко Костянтин Миколайович,
молодший науковий співробітник

Місце роботи автора:
Український мовно-інформаційний фонд НАН України,
вул. Володимирська, 54, м. Київ, 01601, Україна
Тел. (044) 267 4895
E-mail: watcher@enger.kiev.ua