УДК 658.012.011.56

Н.М. СИДОРЧУК

ОРГАНІЗАЦІЯ ДАНИХ ТА ФУНКЦІОНАЛЬНА СТРУКТУРА ЛЕКСИКОГРАФІЧНОЇ СИСТЕМИ «УКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ ЛІНГВІСТИЧНИЙ КОРПУС»

Abstract: *The paper deals with the problems which have emerged in the course of development of the lexicographical system "Ukrainian National Linguistic Corpus" (UNLC). Peculiarities of the metadata storage organization in the subsystem "Digital library" are discussed. An overview of concrete data structures defined for building an access unit to storage objects is presented. The choice of multilevel architecture of the software is substantiated. Functional decomposition of the application level and distinctive features of the data presentation level organization are described.*
Key words: *lexicographical system, linguistic technologies, functional structure, data organization.*

Аннотация: *В работе очерчен круг проблемных задач, возникших во время разработки лексикографической системы «Украинский национальный лингвистический корпус» (УНЛК) и рассмотрены методы их решения. Рассмотрены особенности организации сохранения метаданных подсистемы «Электронная библиотека», и представлен обзор конкретных структур данных, определённых для построения модуля доступа к объектам хранения УНЛК. Обоснован выбор многоуровневой архитектуры программной системы. Представлена функциональная декомпозиция уровня логики приложения и особенности организации уровня представления данных.*
Ключевые слова: *лексикографическая система, лингвистические технологии, функциональная структура, организация данных.*

Анотація: *У роботі окреслено коло проблемних задач, які постали під час розробки лексикографічної системи «Український національний лінгвістичний корпус» (УНЛК) та розглянуто методи їх вирішення. Звернено увагу на особливості організації збереження метаданих підсистеми «Електронна бібліотека» та представлено огляд конкретних структур даних, визначених для побудови модуля доступу до об'єктів збереження УНЛК. Обґрунтовано вибір багаторівневої архітектури програмної системи. Представлено функціональну декомпозицію рівня логіки застосувань та особливості організації рівня представлення даних.*
Ключові слова: *лексикографічна система, лінгвістичні технології, функціональна структура, організація даних.*

1. Вступ

В Українському мовно-інформаційному фонді НАН України ведуться роботи з розробки фундаментальної лексикографічної системи «Український національний лінгвістичний корпус» (УНЛК) [1]. При проектуванні та розробці цієї лексикографічної системи стало очевидно, що, за своєю природою, вона повинна стати інформаційною системою четвертого покоління [2] – мовно-інформаційною системою, яка є інтелектуально орієнтованою і базується на використанні механізмів природної мови.

Застосування новітніх лінгвістичних технологій, ефективне опрацювання великих текстових масивів, паралельне обслуговування великої кількості клієнтів, розподілення функцій системи за різними групами користувачів, забезпечення масштабованості системи, досягнення високого рівня відмовостійкості застосувань, забезпечення надійності збереження та обміну даними, проведення ресурсоємних обчислень та обробки даних, ефективне функціонування розподілених застосувань у глобальній мережі в онлайновому режимі – це далеко не повний перелік тих завдань, що постав при розробці зазначеної системи.

2. Декомпозиція програмного комплексу

В інформаційних проектах такого масштабу вибір генеральної лінії є одним із основних питань, що визначає успіх впровадження та ефективність використання системи. Найпершим аспектом, який слід проаналізувати на предмет адекватності забезпечення якісних властивостей інформаційної

системи, виступає її архітектура. При аналізі поставлених вимог до системи вибір розподіленої архітектури стає очевидним. Така технологія забезпечує централізоване збереження та обробку даних, надає можливість розподіленого введення даних, вирішує проблему обмеження доступу до ресурсів, забезпечує можливість використання потужних обчислювальних можливостей сервера. Програмний комплекс УНЛК реалізовано за трирівневою схемою, у складі якої виділяють рівень даних, логіки застосування та рівень представлення даних [3]. При такій архітектурі програмної системи проміжний рівень (логіки застосувань) перевіряє правильність даних, що передаються від клієнта, та обробляє їх у відповідності з певними правилами. Ця обробка може включати взаємодію з рівнем даних або ж виконувати локальні обчислення чи перетворення, результати яких передаються на рівень даних для збереження, або ж на рівень представлення (клієнтський). Використання такої архітектури дає можливість логічного розподілення функцій системи, що, у свою чергу, забезпечує можливість розподілення роботи між різними розробниками, можливість розробляти окремо кожний рівень, переносити на інші сервери в залежності від вимог масштабованості. Зосередження логіки застосування на проміжному рівні дозволяє модифікувати її, не змінюючи клієнтські системи та інформаційні масиви. І навпаки, з'являється можливість розробки різних клієнтських програм, що використовують один і той же рівень логіки застосувань.

3. Організація збереження метаданих

Основою для розробки будь-якого корпусу повинна бути, перш за все, колекція електронних ресурсів. Метою розробки електронної бібліотеки як компонента УНЛК стало створення спеціального середовища для збору, збереження, моделювання і використання природомовної інформації в цифровому вигляді. Принципи організації даного програмного комплексу повинні були, за задумом розробників, представити можливість створення вхідних потоків лінгвістичної інформації для різноманітних дослідницьких систем, а також забезпечити їх інтеграцію до складу інструментальних засобів електронної бібліотеки. Отже, електронна бібліотека є невід'ємною частиною УНЛК – вона виконує роль багатофункціональної інформаційної системи, яка акумулює інформацію різних видів. У свою чергу, систему „Електронної бібліотеки" можна представити у вигляді декомпозиції на такі елементи: підсистеми збереження об'єктів, підсистеми збереження метаданих та модуля доступу до об'єктів через метадані. За функціональним призначенням система „Електронна бібліотека" покликана забезпечувати реалізацію двох основних завдань: по-перше, інтеграцію в єдиному середовищі інформаційних ресурсів різних типів та видів, а по-друге, забезпечення можливості виокремлення масивів необхідної інформації за заданими критеріями. На наш погляд, ефективність використання електронної бібліотеки можлива лише за умови використання чіткої та прозорої схеми представлення метаданих об'єктів збереження. Питання стандартизації опису даних розглядається як в середині кожної організації, так і на державному і міжнародному рівні. Ми не виключаємо необхідність та актуальність дотримання відповідності установленим стандартам. Та, як переконує практика, при створенні колекцій об'єктів різнорідної природи для різних масивів об'єктів використовується різна підмножина відповідного стандарту - формат опису даних. Звичайно, можна визначити єдиний формат опису всіх об'єктів, але це одразу знизить інтероперабельність системи. Наведемо приклад, що ілюструє цю різнорідність.

В електронній бібліотеці УМІФ НАН України зберігаються як електронні тексти художньої літератури, так і тексти законодавчих актів України. Для опису перших використовуються такі поля, як жанр, стиль, УДК, автор, видавництво, місце видання, ISBN, рік видання і т.д. Для забезпечення зручної пошукової системи відносно текстів законодавчих актів необхідно зберігати таку інформацію, як реєстраційний номер документа, відповідний орган державної влади, дату прийняття документа, стан законодавчого документа. Отже, навіть відносно подібним за природою інформаційним ресурсам (в одному і другому випадку – це текст), відповідає різний формат опису метаданих.

Розглянемо це питання з точки зору системотехніки організації електронної бібліотеки. У випадку, коли внутрішня структура збереження метаданих тісно пов'язана зі встановленим форматом опису даних, ми наштовхуємося на проблему масштабованості системи. Будь-яка зміна формату (навіть незначна) викликає лавиноподібну зміну всього програмного комплексу, що, в свою чергу, призводить до необхідності розробки окремих програмних систем для різних колекцій або зумовлює неперервний процес розробки та супроводження програмної системи. Вирішенням таких питань стала відмова від прив'язки структури бази даних збереження метаданих об'єктів електронної бібліотеки до конкретних даних. Оперуючи поняттями архітектури системи, відповідність даних до сутностей переноситься з рівня логіки роботи з даними на рівень логіки застосувань (рис. 1). Подібний підхід дозволив спроектувати базу даних таким чином, щоб вона слугувала універсальним сховищем збереження метаданих різної природи, а її структура не залежала від формату опису об'єктів збереження.

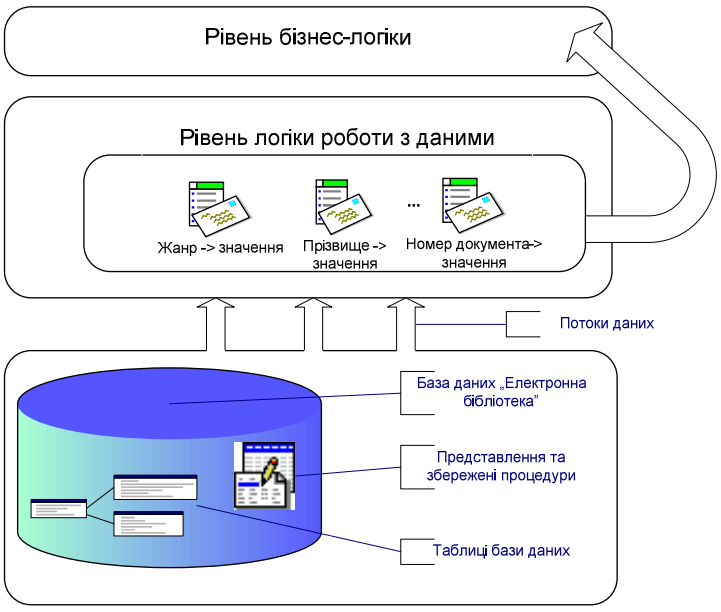

Рис. 1. Зміна підходу організації збереження метаданих

Перш ніж перейти до розгляду структури бази даних, узгодимо деякі поняття:

Об'єкт збереження – це електронний ресурс, який внесений до електронної бібліотеки як цілісна одиниця.

Характеристика – це одиниця опису об'єкта збереження, яка забезпечує можливість його ідентифікації та може використовуватися для пошукових цілей.

Профіль – множина характеристик, якими описуються об'єкти збереження спорідненої природи.

4. Структура бази даних

Зупинимось на розгляді взаємозв'язків між таблицями та структурі кожної з них.

Словники

Поля таблиці:

- Ідентифікатор словника – унікальний в межах системи код.
- Назва словника.
- Опис словника – додаткова інформація про словник (обсяг, призначення, джерело даних і т.д.)
- Закритий чи відкритий. Ознака, яка визначає можливість зміни наповнення даного словника. *True* – множина значень словника може бути змінена користувачем з відповідними правами доступу, *false* – множина значень ініціюється при розробці системи і не може бути змінена.

Словник значень

Поля таблиці:

- Ідентифікатор словника – зовнішній ключ, який реалізує зв'язок з таблицею „Словники”.
- Ідентифікатор в середині словника – для кожного з елементів множини значень словника визначається унікальний ідентифікатор.
- Значення – елемент множини значень словника.
- Опис значення – додаткова інформація до значення зі словника.

Словник характеристик

Поля таблиці:

- Ідентифікатор характеристики.
- Назва характеристики.
- Тип даних. В межах даної системи було визначено п’ять типів даних для характеристики: числовий, символьний, дата та час, шлях до файлу, BLOB.
- Опис характеристики – додаткова інформація про характеристику, яка пояснює специфіку використання даної характеристики, містить посилання на відповідний стандарт.
- Ідентифікатор словника для характеристики. Якщо для характеристики визначений словник значень, то це означає, що дана характеристика може приймати лише значення зі словника.
- Можливість повторення – поле, яке може приймати одне з булівських значень: *true* – для одиниці збереження дана характеристика не може повторюватися; *false* – характеристика може повторюватися.
- Група характеристик (обов’язкова) – одне й те ж значення для декількох характеристик визначає неможливість використання однієї з цих характеристик окремо. Наприклад, об’єкт “художній текст” може бути описаний з використанням таких двох характеристик: мова видання і тип мови видання, об’єднаних в одну групу. Це означає, що внесення до метаданих значення однієї з характеристик вимагає зазначення і другої характеристики. Аналогічно, вилучення з опису однієї з характеристик групи призводить до вилучення всіх інших значень характеристик цієї групи.
- Група характеристик (логічна) дозволяє об’єднати декілька характеристик в логічні групи. На відміну від попереднього поля, характеристики, що об’єднані в логічні групи, можуть існувати окремо одна від одної. Та якщо при описі об’єкта використовуються декілька характеристик з однієї логічної групи, їх послідовність повинна бути чітко визначена.

Користувачі

Таблиця для збереження реєстраційних даних та прав доступу користувачів системи.

Поля таблиці:

- Ідентифікатор користувача – унікальний код користувача в межах системи.
- ПІБ – Прізвище, ім’я та по-батькові користувача системи.
- Псевдонім – ім’я користувача, яке використовується для доступу в комп’ютерну систему (login).
- Тип доступу – ідентифікатор рівня доступу, що визначає привілеї користувача.
- Хеш-функція паролю – поле для збереження закодованого паролю користувача.
- Синхропосилка – значення синхропосилки для проведення поцедури аутентифікації користувача.

Словник профілів

Поля таблиці:

- Ідентифікатор профілю – унікальний код у межах системи.
- Назва профілю.
- Опис профілю – додаткова інформація, що може містити відомості про призначення профілю, дату створення і т.п.

Профілі

Поля таблиці:

- Ідентифікатор профілю – зовнішній ключ, що реалізує зв'язок з таблицею „Словник профілів”.
- Ідентифікатор характеристики – зовнішній ключ, який визначає код характеристики в межах таблиці „Словник характеристик”.
- Упорядкування – визначає порядок слідування характеристик у профілі.

Метадані об’єктів

Поля таблиці:

- Ідентифікатор запису – зовнішній ключ, що реалізує зв'язок з таблицею „Об’єкти ЕлБіб”.
- Ідентифікатор характеристики – зовнішній ключ, що реалізує зв'язок з таблицею „Словник характеристик”.
- Порядок у групі – упорядкування для метаданих одного об’єкта збереження.
- Значення числове – поле, де зберігаються дані для характеристик, в яких визначений тип даних „числовий”.
- Значення символьне – поле, де зберігаються дані для характеристик, в яких визначений тип даних „символьний”.
- Значення дати – дані типу „дата”.
- Значення blob – дані типу „blob”.

Об’єкти збереження

Поля таблиці:

- Ідентифікатор об’єкта збереження – унікальний код об’єкта в межах системи.
- Дата створення.
- Ідентифікатор користувача – зовнішній ключ, що реалізує зв'язок з таблицею „Користувачі” та слугує для збереження інформації про користувача, який відповідає за створений об’єкт.

- Статус об’єкта – ідентифікатор поточного стану об’єкта („новий”, „редагується”, „редагування закінчено”).
- Короткий бібліографічний опис об’єкта – загальні відомості про об’єкт, які формуються на основі метаданих про об’єкт за певними правилами в залежності від природи об’єкта.
- Текст для індексування – поле містить ідентифікатор blob-об’єкта, в якому збережено текст, спеціально підготовлений для проведення індексації.

На рівні логіки роботи з даними реалізовані деякі функції та процедури, що забезпечують коректну роботу з даними та служать для збереження їх цілісності. Так, наприклад, при вилученні характеристики з таблиці „Словник характеристик” постала необхідність реалізації каскадного вилучення всіх записів з таблиць „Профілі” та „Метадані об’єктів”, які посилаються на цю характеристику. З використанням засобів мови програмування PL/pgSQL була написана відповідна функція, що виконує вищезазначену процедуру при кожному застосуванні операції вилучення для таблиці „Словник характеристик”.

5. Програмна платформа розробки системи

Розробка системи УНЛК проводиться на базі платформи .NET [4]. Наявність ієрархічної множини уніфікованих бібліотек класів, потужна та надійна технологія доступу до різних сховищ даних, наявність засобів створення багаторівневих застосувань, забезпечення можливості взаємодії об’єктів, що знаходяться в різних процесах чи доменах застосувань, можливості використання будь-яких мов програмування, які відповідають специфікації CLS, можливості розробки розподілених застосувань для глобальної мережі стали вагомими аргументами при виборі платформи для розробки. За рахунок підтримки таких стандартів, як HTTP, SOAP, WSDL та XML, платформа .Net Remoting дозволяє досягнути максимальної відкритості системи. Для забезпечення максимальної ефективності інфраструктури .NET Remoting у розробників є можливість проводити передачу даних за протоколом TCP.

6. Рівень логіки застосувань

Рівень логіки застосувань становить ядро всього програмного комплексу та реалізує основні серверні функції. На сьогодні в Українському мовно-інформаційному фонді НАН України сервер функціонує під управлінням операційної системи Windows Server 2003 у вигляді сервісу.

Зазначимо, що основні лінгвістичні функції, завдання обробки тексту, підготовка даних до збереження в структурах бази даних, функції моніторингу та адміністрування виконуються на зазначеному рівні. В узагальненому плані їх можна класифікувати на функції електронної бібліотеки та лінгвістичної підсистеми.

Функції електронної бібліотеки:

– формування короткого бібліографічного опису за правилами бібліографування на основі занесених в базу даних елементів метаданих об’єкта збереження;

– формування розгорнутого бібліографічного опису об’єкта збереження;

– редагування множини метаданих бібліографічного опису у відповідності до змін, внесених бібліографом;

– проведення аналізу внесених змін до бібліографічного запису;

– робота з об’єктами файлової системи;

– редагування, вставка, вилучення профілів, характеристик, словників та їх елементів.

Функції лінгвістичної підсистеми:

– створення індексних структур;

– очищення індексних структур;

– індексування об’єктів;

– видалення проіндексованого об’єкта з бази даних повнотекстового індексу;

– повнотекстовий пошук слів та словосполучень у всіх книгах або у книгах, відібраних за бібліографічним описом, з можливістю задавати відстань між пошуковими словами;

– визначення граматичних параметрів слова;

– функції роботи з лексикографічними системами граматичного словника (лематизація, вибір граматичних параметрів, автоматична побудова словозміни і т.п.), словником синонімів (побудова синонімічних рядів, вибір тлумачень) та лексикографічною системою тлумачного словника (побудова розгорнутої структури словникової статті, вибір структурних елементів і т.п.);

– підготовка статистичних даних;

– підготовка мікроконтекстів.

7. Рівень представлення даних

Незважаючи на те, що кожен рівень програмної системи несе своє функціональне навантаження, виконує процедури обміну, обробки та збереження даних і т.п., для кінцевого користувача залишається відкритим лише представницький рівень, і тому питання організації інтерфейсу користувача залишається завжди одним із важливих питань при розробці програмної системи. Інтерфейс користувача реалізовано з підтримкою трьох рівнів доступу:

– користувач (вхід до системи без реєстрації);

– редактор;

– адміністратор системи.

Для роботи з можливістю редактора чи адміністратора проводиться аутентифікація.

Кінцевому користувачеві надаються такі можливості:

– перегляд бібліографічних описів об’єктів, представлених в алфавітному порядку;

– ознайомлення з детальним бібліографічним описом;

– отримання доступу до відповідного об’єкта (тексту, архіву та ін.)

Для зручності користування реалізована пошукова система. Пошук може бути виконаний за такими параметрами: назва видання; прізвище автора (редактора, укладача, колективного автора); стиль; серія; жанр; рік видання (або проміжок часу); анотація; примітки; відомості про видання; місце видання; видавництво; відомості про відповідальність; відомості, що відносяться до назви; ISBN чи ISSN. При чому параметри пошуку можуть бути задані в довільному поєднанні без дотримання відповідності регістру.

Для зручності заповнення бібліографічних реквізитів стосовно видання всю вхідну інформацію було розбито на групи та представлено у вигляді ієрархічного дерева зі зручним

доступом до будь-якого елемента структури. Зауважимо, що реалізовано всі можливі перевірки на коректність заповнюваної інформації:

– виключення дублювання даних;

– можливість вибору із затвердженого переліку;

– відповідність шляху знаходження об’єкта до його опису;

– узгодженість типів;

– правильність написання ISBN (ISSN).

Бібліограф при заповненні інформації у змозі одночасно спостерігати динамічно сформований бібліографічний опис у тому вигляді, в якому він буде доступний безпосередньому користувачеві.

Будучи невід'ємною частиною УНЛК, об’єкт електронної бібліотеки постачає інформацію для лінгвістичної підсистеми. Завдяки інтеграції таких компонентів УНЛК, як електронна бібліотека та лінгвістична підсистема, відпадає необхідність зберігати мікроконтексти (аналоги колишніх лексичних карток) у явному вигляді – для будь-якого слова з реєстру нового словника вони є віртуальними об'єктами і генеруються автоматично на час їх необхідності.

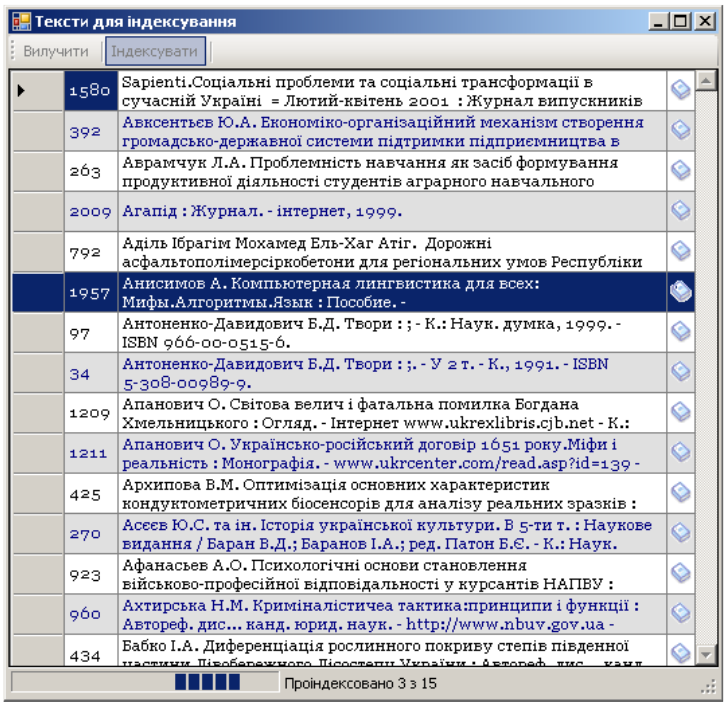

Рис. 2. Створення повнотекстового індексу для обраних об’єктів

Лінгвістична підсистема надає такі функції користувачу:

– створення бази даних повнотекстового індексу;

– очищення бази даних повнотекстового індексу;

– створення черги об’єктів на індексування;

– запуск індексування об’єктів (рис. 5);

– видалення проіндексованого об’єкта з бази даних повнотекстового індексу;

– повнотекстовий пошук слів та словосполучень за варіативними схемами (рис. 6);

– перегляд статистики;

– вибір параметрів створення мікроконтекстів;

– перегляд мікроконтекстів;

– запис мікроконтекстів слів та словосполучень у файл;

– сервісні функції обслуговування.

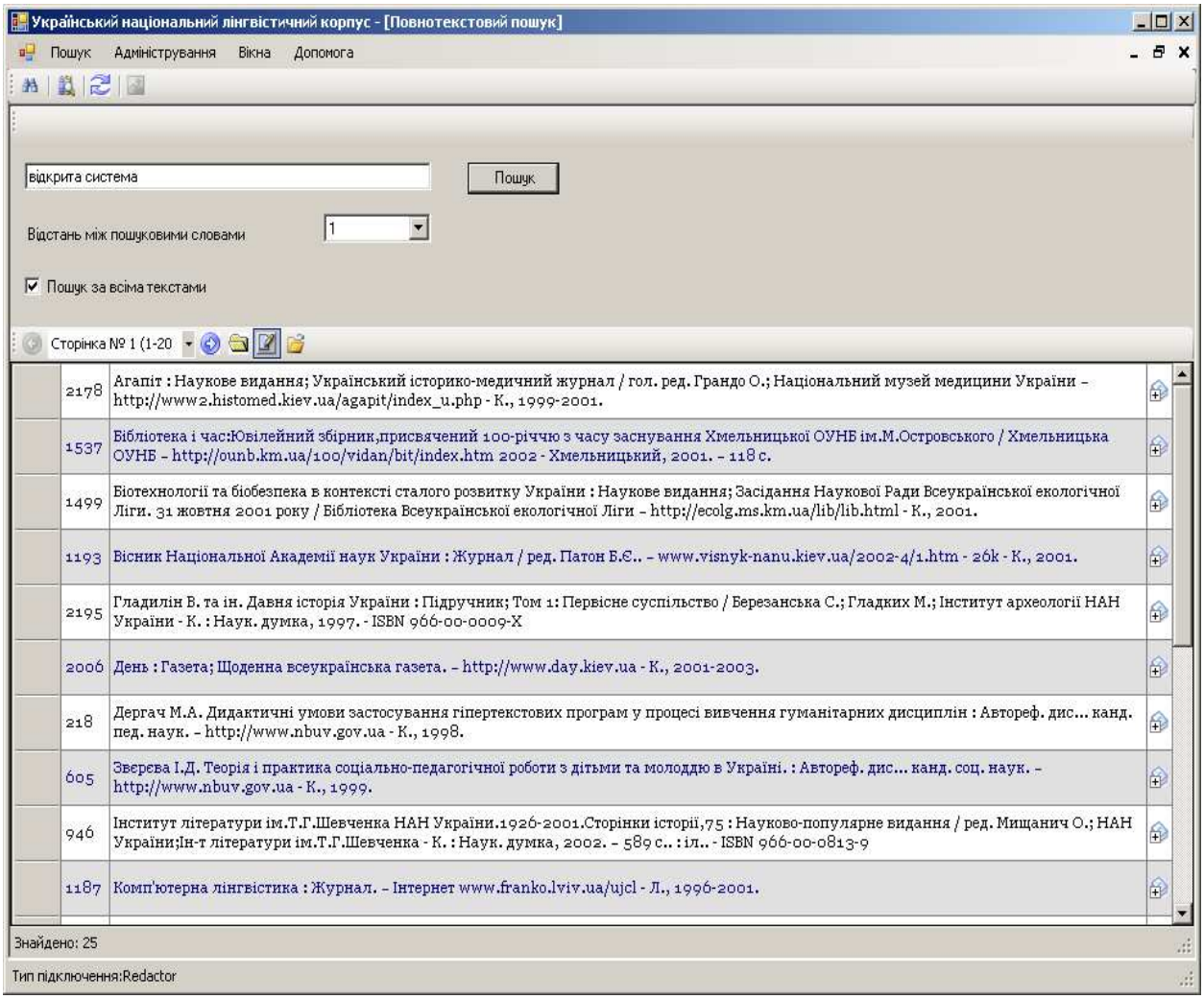

Рис. 3. Інтерфейс повнотекстового пошуку

8. Висновки

Зазначимо, що система, технічні аспекти функціонування якої розкриті в даній статті, функціонує у промисловому режимі в Українському мовно-інформаційному фонді НАН України. Під час експериментальної експлуатації та роботи системи «Українського національного лінгвістичного корпусу» стало очевидно, що питання організації даних, ефективного вибору програмних і технічних засобів, виважений функціональний розподіл за рівнями програмного комплексу має базовий характер при розробці будь-якого мовно-інформаційного проекту великого масштабу. На базі розробленої системи виконуються різноманітні лінгвістичні дослідження, програмний комплекс використовується як додатковий інструмент при розробці нових лексикографічних систем. У подальшому розвиток даного програмного комплексу ми бачимо за такими напрямками:

– розширення та удосконалення лінгвістичного наповнення лексикографічної бази та наповнення електронної бібліотеки;

– інтеграція з іншими лексикографічними системами, зокрема, з тлумачним, етимологічним, термінологічними та іншими словниками української мови;

– реалізація додаткових функціональних можливостей;

– удосконалення інтерактивного спілкування з користувачами;

– можливість використання варіативних інтерфейсних схем та ін.

СПИСОК ЛІТЕРАТУРИ

1. Корпусна лінгвістика / Широков В.А., Бугаков О.В., Грязнухіна Т.О., Костишин О.М., Кригін М.Ю., Любченко Т.П., Рабулець О.Г., Сидоренко О.О., Сидорчук Н.М., Шевченко І.В., Шипнівська О.О., Якименко К.М. – К.: Довіра, 2005. – 471 с.
2. Широков В.А. Елементи лексикографії. – К.: Довіра, 2005. – 304 с.
3. Басс Л., Клементс П., Кайман Р. Архитектура программного обеспечения на практике. – 2-е изд. – СПб.: Питер, 2006. – 575 с.
4. Маклин С. и др. Microsoft. NET Remoting: Пер. с англ. Нафтел Дж., Уильмс К. – М.: Русская редакция, 2003. – 384 с.
5. Широков В.А. Інформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – 331 с.
6. Широков В.А.Феноменологія лексикографічних систем. – К.: Наукова. думка, 2004. – 327 с.