НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ

НАЦІОНАЛЬНА БІБЛІОТЕКА УКРАЇНИ імені В.І.ВЕРНАДСЬКОГО

СИДОРЧУК НАДІЯ МИКОЛАЇВНА

УДК 658.012.011.56

ОНЛАЙНОВІ ЛЕКСИКОГРАФІЧНІ СИСТЕМИ

Спеціальність - 05.13.06 Автоматизовані системи управління

та прогресивні інформаційні технології

Автореферат

дисертації на здобуття наукового ступеня

кандидата технічних наук

Київ - 2006

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Характерною рисою сучасності є активізація інформаційних процесів, потужним інструментом якої стали мережеві інформаційні технології, що відкрили нові напрямки у розвитку і функціонуванні інформаційних систем. Світовою системою комп'ютерних комунікацій щодня користуються сотні мільйонів людей. Інформація, отримана зі всесвітньої мережі, стає одним із визначальних чинників у багатьох галузях народного господарства. Саме вона є продуктом наукової та дослідницької діяльності, необхідним компонентом у ході наукових досліджень та об'єктом практичних застосувань.

Поряд з цим відзначимо, що для розвитку сучасних інформаційних систем стає очевидною орієнтація на використання природномовних механізмів. Зростає потреба у засобах структурування, накопичення, зберігання, пошуку та передачі інформації, створенні інтелектуальних систем обробки інформації та інтелектуальних людино-машинних інтерфейсів. Задоволення саме цих потреб без глибокого розуміння механізмів природної мови взагалі вважається немислимим. Отже, постає **актуальне завдання** розробки загальних методів, технологій та системотехнічних рішень для лексикографічних систем (Л-систем), орієнтованих на застосування Інтернет-технологій та створення на їх основі прикладних україномовних інформаційно-лінгвістичних продуктів, які, по-перше, охоплюють мовні явища в їх цілісності та реальних обсягах, а по-друге – були б адаптовані до використання в ролі елементів онлайнових інформаційних систем, орієнтованих на функціонування в глобальному мережевому середовищі.

Саме такі онлайнові лексикографічні системи спроможні забезпечити адекватний інструментарій при розв'язанні проблем створення інтелектуальних інформаційних систем та розробці інших комп'ютерних засобів опрацювання природної мови, орієнтованих на функціонування в сучасних мережевих середовищах.

Розв'язанню цієї проблеми і присвячена наша дисертаційна робота **"Онлайнові лексикографічні системи"**.

Зв'язок дослідження з науковими програмами, планами, темами.

Результати дисертаційного дослідження одержано в процесі виконання загальнодержавної та відомчої тематики, а саме завдань, визначених:

1. Розпорядженням Президії НАН України від 06.06.2005 № 350 «Про затвердження переліку проектів програми інформатизації Національної академії наук України на 2005 рік» (проект «Розробка інструментальної системи створення та ведення цифрового архіву документів Президії НАН України», № держреєстрації НДР 0105Г008376).

2. Розпорядженням Президії НАН України від 19.05.06 № 301 «Про затвердження переліку проектів комплексної програми наукових досліджень «Соціально-економічні і гуманітарні

чинники інноваційного розвитку України» (проект «Розробка інструментальної системи електронного Українського біографічного архіву», № держреєстрації НДР 0104U008703)

3. Розпорядженням Президії НАН України від 19.05.06 № 301 «Про затвердження переліку проектів комплексної програми наукових досліджень «Соціально-економічні і гуманітарні чинники інноваційного розвитку України» (проект «Розробка теоретико-методологічних і системотехнічних засад Національного депозитарію лінгвістичних ресурсів України», № держреєстрації НДР 0104U008704).

Мета і завдання дослідження. Метою дисертаційної роботи є розробка теорії онлайнових лексикографічних систем, загальних принципів їх функціонування та реалізація конкретних онлайнових програмно-лінгвістичних комплексів на базі отриманих результатів.

Досягнення цієї мети передбачає виконання таких **науково-технічних завдань**:

- конкретизація теорії лексикографічних систем на випадок функціонування їх в мережевому середовищі (глобальному та локальному);
- розробка моделей онлайнових лексикографічних систем («Інтегрованої Л-системи «Словники України», Українського лінгвістичного порталу, електронної бібліотеки з розширеними лінгвістичними функціями, Українського національного лінгвістичного корпусу, Українського біографічного архіву тощо);
- дослідження загальних принципів створення та функціонування онлайнових лексикографічних систем;
- створення лексикографічних баз даних для розроблених онлайнових моделей лексикографічних систем;
- реалізація зазначених лексикографічних систем у мережевих середовищах.

Об’єктом дослідження є онлайнові лексикографічні системи.

Предметом дослідження є взаємодія архітектур відкритих систем в інтерпретації OSI-7 та архітектури лексикографічних систем в ANSI/X3/SPARK інтерпретації.

Методи дослідження ґрунтуються на комп’ютерному моделюванні мовно-інформаційних процесів у мережевих комп’ютерних середовищах. Використовується теорія лексикографічних систем, теорія моделей та баз даних, формалізована інтерпретація мовного матеріалу, архітектури відкритих систем і лексикографічних систем; сучасні методи та інструментальні засоби програмування.

Наукова новизна. Розроблено принципи взаємодії архітектури OSI-7 та архітектури лексикографічних систем в ANSI/X3/SPARK інтерпретації. Сформульовано системотехнічні засади розробки онлайнових інформаційно-лексикографічних систем та баз даних, які забезпечують принципи платформонезалежності, масштабованості, безпеки, практичності та продуктивності. Зазначені засади

апробовано на прикладах реальних, промислово діючих варіантів онлайнових лексикографічних систем (Український лінгвістичний портал тощо).

Практичне значення отриманих результатів. Створено низку онлайнових інформаційно-лексикографічних систем, що широко використовуються відвідувачами з десятків країн світу (Український лінгвістичний портал: http://ulif.org.ua, яку впроваджено в Міністерстві освіти та науки України: http://ulif.mon.gov.ua). Розроблено низку інструментальних інформаційно-лінгвістичних систем для розв'язання актуальних завдань Національної академії наук України (Український біографічний архів, Архів документів Президії НАН України, варіант Інтегрованої Л-системи «Словники України» для глобальних мереж). Програмне забезпечення мережевої цифрової бібліотеки впроваджено в Уманському державному педагогічному університеті імені Павла Тичини. Розроблено системотехніку Українського національного лінгвістичного корпусу (УНЛК) з розширеними лінгвістичними функціями, адаптовану до роботи у мережевому середовищі.

Апробація результатів дисертації. Основні положення та результати дисертаційного дослідження були висвітлені на міжнародних конференціях: «Корпусна лінгвістика 2004» (Санкт-Петербург, 2004 р.), «Інформаційні системи та технології» в рамках 2-го Міжнародного радіоелектронного форуму «Прикладна радіоелектроніка» (Харків, 2005), «Інформація для всіх: культура та технології Інформаційного суспільства» (Москва, 2005), "Інтелектуальні інформаційні технології у бібліотечній справі" (Київ, 2005), «Горизонти прикладної лінгвістики та лінгвістичних технологій» (Крим, Партеніт, 2006) та численних семінарах Українського мовно-інформаційного фонду НАН України.

Публікації з теми дисертації. Результати роботи висвітлені у десяти наукових працях, три з яких – одноосібні статті у фахових журналах, що входять до переліку ВАК України, одна - колективна монографія «Корпусна лінгвістика», п'ять – тези та праці міжнародних наукових конференцій, одна – електронна публікація. Результати дисертації, які винесено на захист, належать авторові.

Особистий внесок здобувача. У публікаціях, написаних у співавторстві, нам належать усі системотехнічні та програмні реалізації. У колективній монографії «Корпусна лінгвістика» автором написана частина другого розділу (разом із О.М.Костишиним; особистий внесок автора – 50%) та частина шостого розділу (разом із О.Г.Рабульцем, К.М.Якименком та О.М. Костишиним; особистий внесок автора – 50%).

Структура роботи. Дисертація складається зі вступу, трьох розділів основного змісту, висновків, чотирьох додатків та списку використаної літератури зі 149 найменувань. Обсяг дисертації без списку використаної літератури – 157 сторінок, загальний обсяг роботи (з бібліографією та додатками) – 208 сторінок.

ОСНОВНИЙ ЗМІСТ РОБОТИ

Перший розділ **«Концептуальне моделювання лексикографічних систем»** присвячено викладу загальної теорії лексикографічних систем та побудові концептуальних моделей лексикографічних систем, які в наступних розділах підлягають системотехнічному моделюванню та програмній реалізації. На основі феноменологічного принципу – лексикографічного ефекту в інформаційних системах – будується загальна схема лексикографічної моделі даних.

Відповідно до інформаційної інтерпретації процесів сприйняття визначається результат рецепції суб'єктом S класу елементарних інформаційних одиниць (ЕІО) $I^{Q}(D)$, де символом D позначено фрагмент дійсності, який перебуває у полі уваги суб'єкта S, а через Q – лексикографічний ефект, який породжує клас $I^{Q}(D)$, у вигляді певної множини $V(I^{Q}(D))$ — множини описів одиниць, що належать до класу $I^{Q}(D)$; ця множина є результатом процесу:

$$S : I^{Q}(D) \rightarrow V(I^{Q}(D)), \tag{1}$$

тому для кожного елемента $x \in I^{Q}(D)$ однозначно визначено його опис $V(x)$ як елемент множини $V(I^{Q}(D))$: $V(x) \in V(I^{Q}(D))$; $Sx = V(x)$. Отже, логічно припустити, що $V(I^{Q}(D))$ має вигляд об'єднання:

$$V(I^{Q}(D)) = \bigcup_{x \in I^{Q}(D)} V(x). \tag{2}$$

Згідно з інформаційною концепцією представлення опису системи ЕІО, кожний $V(x)$ зображується у вигляді слова (тексту) в певному алфавіті $\boldsymbol{A} = \{a_1, a_2, ..., a_n\}$, тобто скінченної послідовності символів з $\boldsymbol{A}$. Надалі слова в алфавіті $\boldsymbol{A}$ називатимемо $\boldsymbol{A}$-словами. На множині (2) індукується структура у такий спосіб. Припустимо, що для всіх описів $V(x)$ існує єдине правило, за яким із будь-якого $\boldsymbol{A}$-слова $V(x)$ можна виділити множину $\boldsymbol{A}$-підслів $\beta(x) = \{\beta_i(x)\}$ із такими властивостями:

– елемент x належить до множини $\beta(x)$;

– весь опис $V(x)$ є елементом множини $\beta(x)$;

– правило, за яким виділяються елементи множини $\beta(x)$ є єдиним для всіх $V(x)$.

Описаним способом із будь-якого $V(x)$ виділяється множина $\beta[V(x)]$ величин ($\boldsymbol{A}$-підслів) $\beta_i(x)$ такого вигляду:

$$\beta[V(x)] \equiv \{\beta_i(x), i = 1, 2, ..., q\} \subseteq B[V(x)], \tag{3}$$

де $B[V(x)] = \{v_{i_1} v_{i_2} ... v_{i_p}$, $1 \le i_1 < i_2 < ... < i_p \le k(x)$, $p = 1, 2, ..., k(x)$, причому:

$$v_{ij} \in \{v_{1(x)}, v_{2(x)}, ..., v_{k(x)}(x)\}; x \in \beta[V(x)]; V(x) \in \beta[V(x)], \beta_{k(x)} \neq \beta_{m(x)} \text{ при } k \neq m. \tag{4}$$

Покладемо за визначенням:

$$\beta[V(I^{Q}(D))] = \bigcup_{x \in I^{Q(D)}} \beta[V(x)]. \tag{5}$$

Очевидно, що $V(I^{Q}(D)) \in \beta[V(I^{Q}(D))]$. Позначимо

$$\beta_i = \bigcup_{x \in I^{Q(D)}}^{q} \beta_i(x),\ i = 1, 2, ..., q, \text{ а також } \beta = \bigcup_{i} \beta_i . \qquad (6)$$

Зрозуміло, що $\beta \equiv \beta\,[V(I^{Q}(D))]$. Деякі з елементів $\beta_i(x)$, $i = 1, 2, ..., q$, можуть бути порожніми для певних $x \in I^{Q}(D)$; у цьому випадку вони випускаються у формулах (3) – (6).

Через $\boldsymbol{\sigma}[\beta]$ позначимо певну структуру, визначену на β і, отже, на $V(I^{Q}(D))$,— надалі називатимемо $\boldsymbol{\sigma}[\beta]$ макроструктурою $V(I^{Q}(D))$; обмеження $\boldsymbol{\sigma}[\beta]$ на $V(x)$: $\boldsymbol{\sigma}[\beta]\,\big|_{V(x)} \equiv \boldsymbol{\sigma}(x)$ породжує мікроструктуру $V(x)$. Активне формулювання цього факту полягає у встановленні процедури (оператора, процесу...) $\boldsymbol{\sigma}$, який породжує на β структуру $\boldsymbol{\sigma}[\beta]$:

$$\boldsymbol{\sigma}: \beta \rightarrow \boldsymbol{\sigma}[\beta]. \qquad (7)$$

На β можлива генерація цілої низки неізоморфних структур $\boldsymbol{\sigma}[\beta]$. У їх ролі можуть виступати будь-які з відомих моделей даних (ієрархічна, мережева, реляційна, об’єктно-реляційна та ін.), логіко-математичні моделі тощо. Отже формулами (3) – (7) визначено певну модель даних, яка має досить загальну природу. Зазначена модель одержує інтерпретацію як лексикографічна система в архітектурі ANSI/X3/SPARK. Описується процес рекурсивної редукції лексикографічної системи та будується конструкція лексикографічного середовища, як основного об’єкта, що використовується при розробці процесів інтеграції Л-систем та побудові інтегрованих Л-систем. Будуються концептуальні моделі Л-систем, що відповідають Інтегрованій Л-системі «Словники України», лінгвістичному корпусу, складовою якого є цифрова бібліотека тощо. У наступних розділах зазначені концептуальні моделі реалізуються як різнопланові онлайнові Л-системи.

У другому розділі «**Системотехнічні засади онлайнових лексикографічних систем»** розглянуто такі питання. Для створення ефективної технології онлайнових лексикографічних систем необхідно узгодити власне архітектуру Л-систем з OSI-архітектурою. Причому зазначене узгодження в основному стосується 5, 6 та 7 рівнів OSI-архітектури та внутрішнього й зовнішнього рівнів архітектури Л-систем в інтерпретації ANSI/X3/SPARK. Л-системи взаємодіють з іншими Л-системами, створюючи тим самим лексикографічне середовище, причому взаємодія відбувається між об’єктами одного й того ж рівня. Таким чином, кожен із визначених рівнів забезпечує сервіс суміжному з ним верхньому рівню; отримує сервіс від суміжного з ним нижнього рівня; здійснює обмін блоками даних з метою виконання певних задач. Визначено критерії якості для онлайнових лексикографічних систем:

- масштабованість; онлайнова система повинна передбачати можливість коливань навантаження та реагувати на них без втручання людини;
- надійність; періоди простою системи повинні зводитися до мінімуму;

- безпека; онлайнова система повинна проводити автентифікацію користувачів запобігаючи несанкціонованому доступу до даних;
- практичність; у різних користувачів повинна бути можливість доступу до різного вмісту у різних формах;
- продуктивність; системи, з якими взаємодіють користувачі, повинні демонструвати високу реактивність.

Визначено архітектуру програмного забезпечення для систем подібного типу. Це багаторівнева архітектура, використання якої дає можливість логічного розподілення функцій системи, що в свою чергу забезпечує можливість розподілення роботи між різними розробниками, можливість розробляти окремо кожен 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 застосуваннння.

Реалізація продуктивності для онлайнових лексикографічних систем полягає в реагуванні на повідомлення, що поступають в систему протягом певного часу. Після надходження повідомлення система розпочинає його обробку. За тих чи інших причин обробка може бути заблокованою. Звідси можна зробити висновок про дві складові часу реакції системи, якими є використання ресурсів і тривалість блокування. Ресурси – це і центральний процесор, і сховища даних, і пропускна здатність мережевих з'єднань, пам'ять тощо. Тривалість блокування залежить від змагання за ресурс, його готовності або залежності даного обчислення від результатів інших обчислень, які ще не завершені. Складність та особливість взаємодії онлайнових систем полягає у тому, що вони повинні мати можливість досягнення високого ступеня незалежності програмних інтерфейсів від великої кількості програмних середовищ різних типів. Багаторівнева архітектура для програмного забезпечення систем подібного типу є як правило базовою, проте особливості предметної галузі вимагають застосування архітектури, що надає ширші можливості у функціонуванні, використанні лексикографічних систем та відтворюванні лексикографічних середовищ. Саме для онлайнових лексикографічних систем як мережевих розподілених застосувань, з огляду на проведене дослідження, сервіс-орієнтована архітектура (Service-Oriented Architecture, SOA) виявляється найбільш доцільною. Взаємодія Л-систем за сервіс-орієнтованою архітектурою ґрунтується на основі чотирьох базових стандартів: розширювана мова розмітки (XML), простий протокол доступу до об'єкта (SOAP), мова опису веб-сервісу (WSDL), універсальний метод опису, виявлення та інтеграції сервісів (UDDI).

Концепція проектування програмного забезпечення як сервісу, незалежного від інтерфейсу користувача, стала базовою для онлайнових лексикографічних систем. Веб-сервіс як

одна з логічних абстракцій сервіс-орієнтованої архітектури забезпечує пошук, опис та ініціювання засобів його роботи простими та прозорими для клієнта методами та функціонує на будь-якій машині, мові та платформі.

Клієнт зв'язується з відкритим загальнодоступним реєстром для пошуку конкретного сервісу. Причому, вузол може містити більше одного сервісу. Так само як і сервіс може бути представлений декількома джерелами. Отримавши сигнатуру сервісу у разі позитивного результату пошуку, клієнт може ініціювати його роботу, оперуючи даними, отриманими у формі документу WSDL. Після ідентифікації сервісу, можливо проведення певних процедур за правилами, що представлені в описі веб-сервісу. Загальна схема взаємодії онлайнових Л-систем представлено на рис.1.

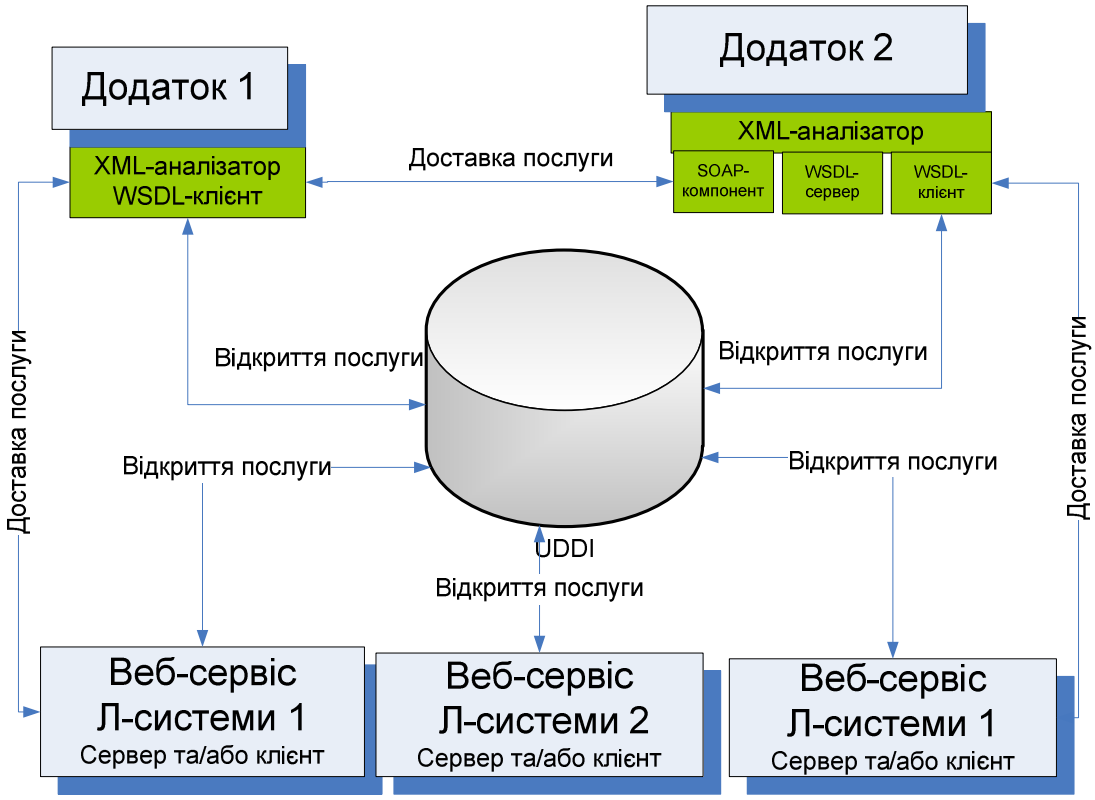

Рис 1. Схема взаємодії онлайнових лексикографічних систем

У третьому розділі **«Приклади побудови онлайнових лексикографічних систем»** розглянуто прикладні аспекти реалізації онлайнових лексикографічних систем, які було спроектовано, розроблено та передано у промислову експлуатацію. Розробка програмних комплексів здійснювалась на основі концептуальних засад, поданих у першому розділі, та базуючись на основних системотехнічних принципах, розроблених та висвітлених у другому розділі. Матеріал подається на прикладі розробки онлайнової лексикографічної системи «Український лінгвістичний портал», результатах, отриманих при реалізації інструментального комплексу «Цифровий архів документів Національної академії наук України», Українського

національного лінгвістичного корпусу та електронної бібліотеки в мережевому середовищі як його невід'ємного компонента.

Український лінгвістичний портал (УЛП) виступає головною точкою входу до лінгвістичних ресурсів, розміщених Українським мовно-інформаційним фондом НАН України в мережі Інтернет. Зазначена Л-система була розроблена та впроваджена в дослідну експлуатацію з червня 2004 року, а з липня 2005 року лексикографічну систему було передано та впроваджено як дзеркало лінгвістичного порталу в Міністерстві освіти та науки України, що підтверджено відповідним актом, який подається у додатку Г до дисертації. УЛП побудовано за багаторівневою технологією клієнт-сервер з використанням таких програмних засобів для розробки: сервер застосувань розроблено на мові інтерпретації сценаріїв PHP4, 5 у поєднанні з HTML та JavaScript сценаріями, система керування базами даних PostgreSQL. До складу Українського лінгвістичного порталу входять три базові онлайнові лексикографічні системи:

- «Словники України он-лайн»;
- «Словник російської словозміни он-лайн»;
- «Онлайновий журнал «Мовознавство».

«Словники України он-лайн» відтворюють у мережевому середовищі інтегрований лексикографічний комплекс „Словники України" з притаманними йому функціями словозміни, синонімії, антонімії та фразеологічною підсистемою. Технологічне ядро зосереджено у спеціальному програмному комплексі, що функціонує в локальній мережі Українського мовно-інформаційного фонду НАН України, тому вирішується завдання підготовки даних для наповнення бази онлайнової лексикографічної системи. Так само, як і в локальній версії, множина входів до системи не обмежується реєстровим рядом, а охоплює і праві частини словникових статей. Кожне слово правої частини є активним – воно проіндексоване і стає додатковою точкою входу до відповідної словникової одиниці. Це свідчить про густу мережу зв'язків у системі, що відкриває великі функціональні можливості при досить простому й прозорому інтерфейсному відображенні та забезпечує високий ступінь інтерактивності. Інформаційні потоки всередині системи, організація інтерфейсу користувача, проблемні аспекти реалізації онлайнового програмного комплексу «Словники України» розглянуто в межах даного дисертаційного дослідження.

Загальний вигляд інтерфейсу представлено на рис. 2, де виділено такі окремі його елементи:

1. Елемент введення, який призначено для пошуку слів у реєстрі словника. Користувач має змогу швидко переміститися на слово, яке його цікавить, або ж, у разі його відсутності у множині реєстрового ряду, на групу найближчих за написанням лексем.

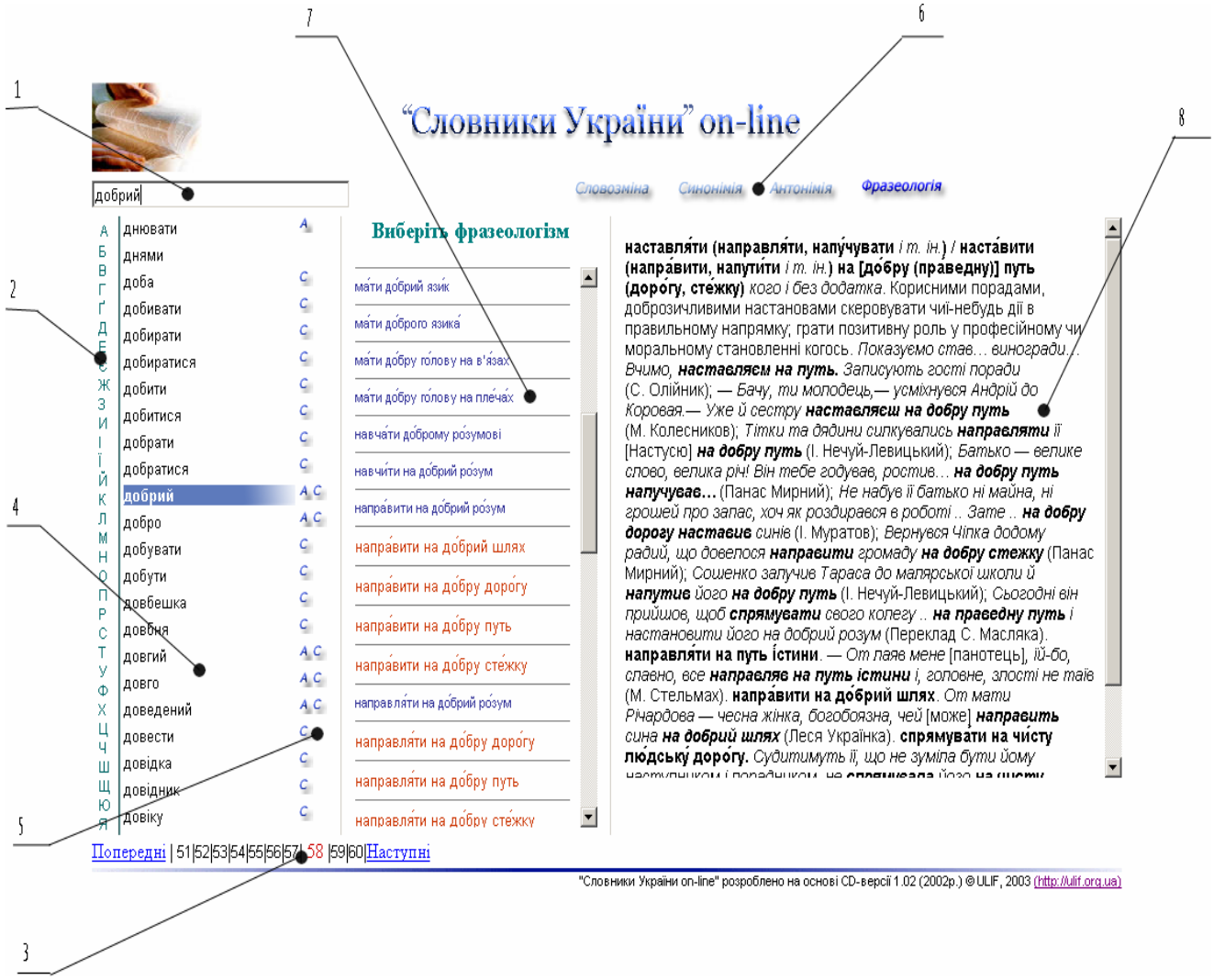

Рис.2. Загальний інтерфейс користувача онлайнової версії інтегрованої лексикографічної системи «Словники України»

2. Додаткова навігація реєстром здійснюється швидким переходом за абетковою літерою. Реалізовано ефект закладок, за допомогою яких відбувається позиціонування на перше слово, що розпочинається з обраної літери.

3. Ще один елемент навігації. Весь реєстровий ряд розбито на так звані віртуальні групи (кількість одиниць в групі задається програмно).

4. Активна частина реєстрового ряду. Обране слово позначено з використанням специфічного стильового відображення.

5. Позначки наявності слова в тому чи іншому словнику: а, с, ф – відповідно в словнику антонімів, синонімів та фразеологічному.

6. Зміна типу словника. Активний словник позначено більш інтенсивним кольоровим забарвленням.

7. Область виведення словникової статті. Реалізована з використанням плаваючого фрейму. Такий підхід дав змогу позбутися проблеми позиціонування при різних величинах правої частини словникових статей. Ще більш виправдано вибір фреймової репрезентації у фразеологічному словнику, де ця область розбивається на дві частини (в першій відображається множина фразеологізмів, що відповідають обраному реєстровому слову, а в другій – інтерпретаційна частина відповідної фразеологічної групи). Саме таке візуальне представлення дозволяє відчути ієрархію зв'язків цього словника.

Одним із об'єктів моделювання онлайнових систем є інструментальна система створення та ведення цифрового архіву документів Президії Національної академії наук України (ІС ЦАД). Ця система призначена для переведення паперового архіву документів Національної академії наук України до цифрової форми із забезпеченням доступу до цих документів користувачам локальної мережі Президії НАН України та інститутів й установ НАН України через мережу Інтернет у режимі он-лайн. ІС ЦАД дозволяє у напівавтоматичному діалоговому режимі створювати електронні версії архівних документів, інформація у яких подана у формі цифрових даних (текстів документів та їхніх цифрових копій). Проектування та реалізація концептуальної моделі метаданих проведені у відповідності до стандарту UniMARC-21, з урахуванням призначення інформаційної системи цифрового архіву Президії НАН України та специфіки її використання. Вдалося отримати такі результати: Цифровий архів Президії НАН України описаної архітектури може зберігати не тільки бібліографічні описи документів та інших об'єктів зберігання (зображень, звукових та відеоматеріалів), а й самі об'єкти; завдяки багаторівневій клієнт-серверній архітектурі цифрового архіву підвищено його масштабованість, значно зменшено завантаження мережі, що є особливо важливою умовою при використанні подібної системи у мережі Інтернет (наприклад, під'єднання інститутів та інших установ Національної академії наук України до архіву); завдяки своїй платформонезалежній архітектурі, цифровий архів природним чином з мінімальними витратами інтегрується як до мережі Інтернет, так і до будь яких корпоративних інформаційних середовищ.

Ще одним онлайновим програмним комплексом є електронна бібліотека, метою розробки якої стало створення спеціального середовища для збору, збереження, моделювання та використання природномовної інформації в цифровому вигляді. В цьому комплексі реалізована багатомовна підтримка, пошуковий механізм за бібліографічними описами, загальні функції читача та автоматизоване робоче місце бібліографа для внесення та редагування необхідної інформації. Для відтворення даної системи в мережевому середовищі була проведена попередня робота зі створення конвертора даних підсистеми бібліографічних описів, які були сформовані раніше засобами CDS/ISIS, в середовище PostgreSQL, а також приєднання їх до відповідних компонентів об'єктів збереження. Структура бази для збереження метаданих наведена на рис. 3.

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.

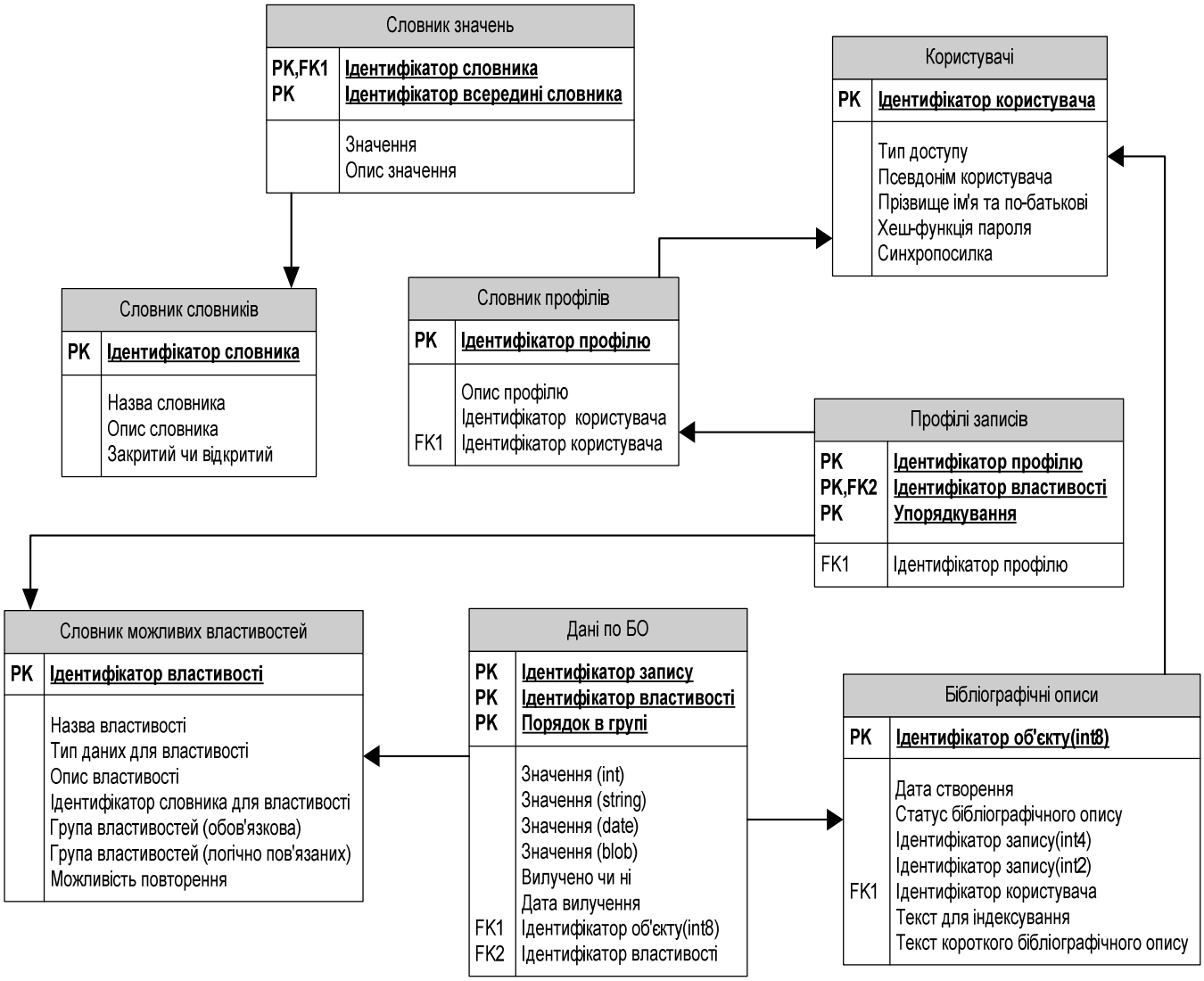

Рис. 3. Схема бази даних „Електронна бібліотека”

Зазначимо, що основні лінгвістичні функції, завдання обробки тексту, підготовка даних до збереження в структури бази даних, функції моніторингу та адміністрування виконуються на рівні логіки застосування. До загальних функцій електронної бібліотеки віднесемо: 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.

З використанням системотехніки електронної бібліотеки було створено нову, вдосконалену версію Українського національного лінгвістичного корпусу. В цій версії вхідною підсистемою є електронна бібліотека. Крім цього, в системі забезпечено функцію автоматичної побудови повного індексу за лексичною системою, що надає можливість для створення пошукового індексу за словами та словосполученнями та виділення будь-якого мікроконтексту, що містить виділену сукупність лексичних одиниць у будь-яких граматичних значеннях.

Як базова 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) в явному вигляді – для будь-якого слова з реєстру нового словника вони є віртуальними об'єктами і генеруються автоматично на час, поки в них є потреба.

На рис. 4 передано інтерфейс головної сторінки Українського національного лінгвістичного корпусу. Доступ до об’єктів збереження в корпусі здійснюється через підсистему пошуку. Цифрами позначено:

1. Пошук за бібліографічними реквізитами, що здійснюється в межах електронної бібліотеки за метаінформацією до об’єктів.
2. Пошук за лінгвістичними параметрами, що реалізується за рахунок повнотекстового індексу.
3. Рядок введення слова або фрази для пошуку в повних текстах документів корпусу.
4. Додаткові параметри повнотекстового пошуку:
   - урахування порядку слів;
   - пошук у певній підмножині об’єктів;
   - з використанням процедури лематизації (зведення кожного пошукового слова до вихідної форми);
   - використання синонімічної лексикографічної бази даних;
   - множина певних синонімічних рядів;
   - множина граматичних параметрів для кожного слова, що входить в пошукову фразу.
5. Користувач має доступ до реєстру граматичного словника (приблизно 230 тис. реєстрових одиниць).
6. Результати проведеного пошуку – множина коротких описів документів, що відповідають заданим критеріям.
7. Функції роботи з поточними документами (перегляд текстів, імпорт результатів у певні формати даних).
8. Перегляд бібліографічних реквізитів.
9. Статистичні дані згідно з індексом за поточним документом.

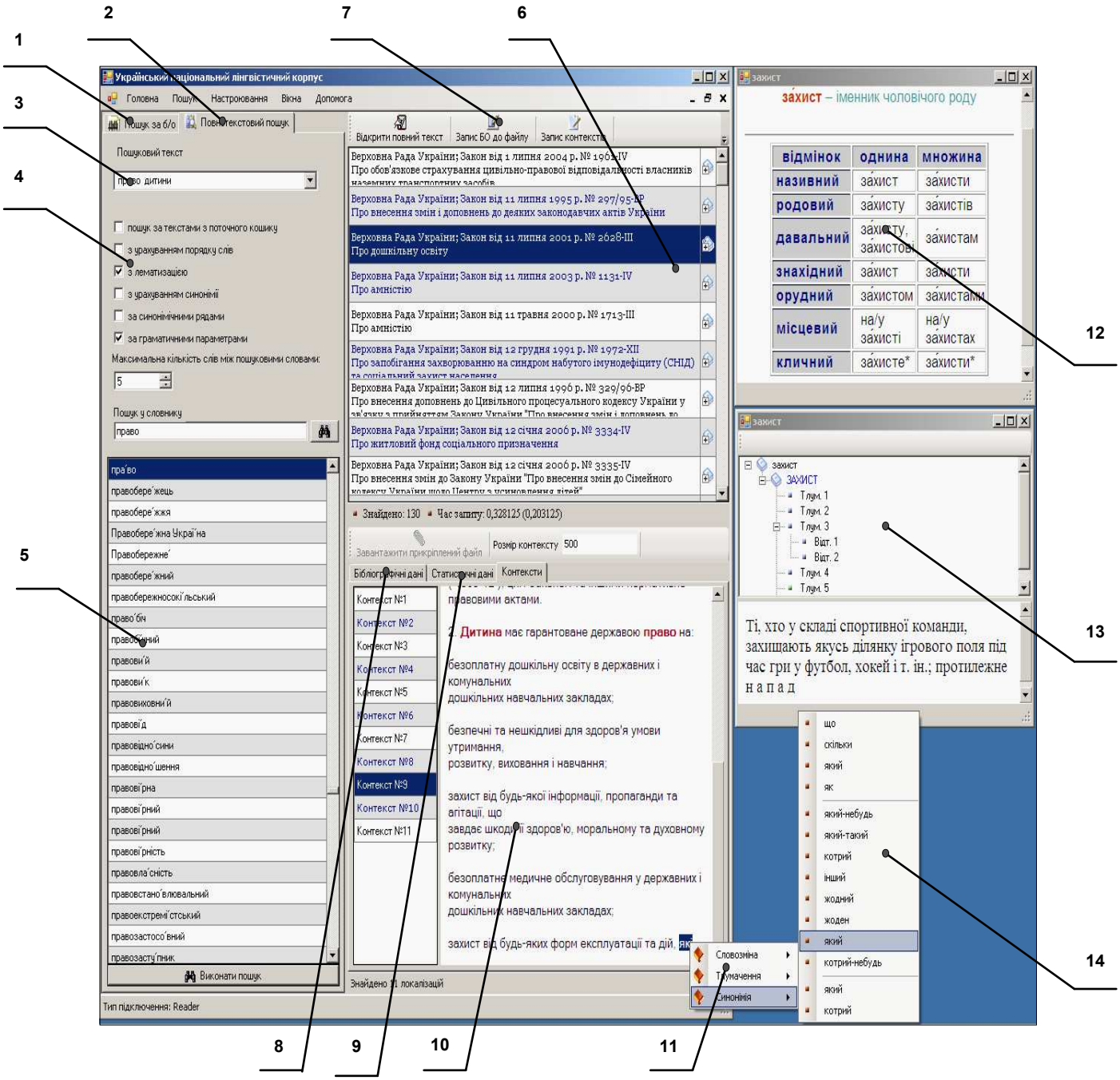

Рис 4. Інтерфейс повнотекстового пошуку УНЛК

10. Множина контекстів, що відповідають заданим критеріям.

11. Для кожного слова реалізовано зв'язок з парадигматичним, синонімічним та тлумачним словниками.

12. Перегляд словозміни для обраного слова.

13. Перегляд тлумачної словникової статті.

14. Розподіл за синонімічними рядами.

ВИСНОВКИ

У ході дисертаційного дослідження розв’язано ряд актуальних науково-технічних проблем української лінгвістичної технології та отримано низку нових наукових та практично цінних результатів.

1. Здійснено конкретизацію теорії лексикографічних систем на випадок їх функціонування в мережевому середовищі (глобальному та локальному).

2. Розроблено моделі онлайнових лексикографічних систем (Інтегрованої Л-системи «Словники України», Українського лінгвістичного порталу, електронної бібліотеки з розширеними лінгвістичними функціями, Українського національного лінгвістичного корпусу тощо).

3. Виконано дослідження програмних архітектур для мережевих застосувань та обґрунтовано доцільність використання сервіс-орієнтованої архітектури для реалізації онлайнових Л-систем. Уперше розроблено принципи взаємодії архітектури OSI-7 та архітектури лексикографічних систем в ANSI/X3/SPARK інтерпретації. Досліджено чотири базові стандарти, на яких ґрунтується інтеграція інформаційних систем на основі веб-сервісів: розширювана мова розмітки, простий протокол доступу до об'єкту, мова опису веб-сервісу, універсальний метод опису, виявлення та інтеграції сервісів. На основі зазначеного дослідження розроблено системотехнічні засади реалізації лексикографічних систем у глобальних мережевих середовищах .

4. Створено лексикографічні бази даних для розроблених моделей онлайнових лексикографічних систем.

5. Розроблено онлайнову Л-систему «Український лінгвістичний портал», яку впроваджено в Міністерстві освіти та науки України. У складі УЛП уперше реалізовано повномасштабну українську інтегровану лексикографічну систему «Словники України» та онлайнову систему «Словник російської словозміни». Створено програмне забезпечення мережених інформаційних систем «Цифровий архів документів Президії НАН України». Розроблено мережевий варіант електронної бібліотеки та Українського національного лінгвістичного корпусу.

ПУБЛІКАЦІЇ З ТЕМИ ДИСЕРТАЦІЇ

1. Широков В.А. та ін. Корпусна лінгвістика: Монографія / 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.; Український мовно-інформаційний фонд НАН України - К. : Довіра, 2005. – 471 с.
2. Сидорчук Н.М. Технологічні аспекти реалізації онлайнової лексикографічної системи „Словники України” //Проблеми програмування. – 2005. – № 4. – С. 95–105.

3. Сидорчук Н.М. Архітектурні та системотехнічні підходи до конструювання Українського національного лінгвістичного корпусу // Бионика интеллекта. – 2005. – № 2(63) . – С. 107-110.
4. Сидорчук Н.М. Організація даних та функціональна структура лексикографічної системи «Український національний лінгвістичний корпус» // Математичні машини і системи. – 2006. – № 2 . – С. 126-135.
5. Рабулец А.Г., Костышин А.М., Сидорчук Н.Н., Широков В.А., Якименко К.Н. Системотехнические и лингвистические принципы украинского лингвистического корпуса // Труды международной конференции «Корпусная лингвистика 2004». – СПб.: Изд-во С.-Петерб. Ун-та, 2004. – С. 286-303.
6. Костишин О.М., Сидорчук Н.М. Системотехнічні та лінгвістичні принципи проектування українського лінгвістичного корпусу // Наукові праці Національної бібліотеки України ім. В.І Вернадського НАН України. Випуск 141. / Нац. Бібліотека України ім. В.І. Вернадського; Редкол .: Онищенко О.С. (гол.) та ін. – К., 2005 . – С. 190-199.
7. Костишин О.М., Сидорчук Н.М. Інструментальна система для створення онлайнових електронних журналів // Міжнародна наукова конференція "Інтелектуальні інформаційні технології у бібліотечній справі": Семінар «Нові інформаційні технології електронних бібліотек» // http://www.nbuv.gov.ua/new/05_Kiev/05komoez.html.
8. Сидорчук Н.М. Онлайнові лексикографічні системи (на прикладі системи „Словники України”) // 2-й Международный радиоэлектронный форум «Прикладная радиоэлектроника. Состояние и перспективы развития» МРФ-2005. Т.3. Международная конференция «Информационные системы и технологии»: Сб. научных трудов. – Х.: АНПРЭ, ХНУРЭ. 2005. – С. 39-42.
9. Остапова И.В., Сидорчук Н.Н. Интернет-версия научного журнала (на примере журнала «Мовознавство» Института языкознания им. А. А. Потебни и Украинского языково-информационного фонда Национальной академии наук Украины) // EVA 2005, Технология информационного общества и культура : Материалы 8-ой ежегодной Международной конференции. – М., Центр ПИК, 2005. – Эл. изд.
10. Широков В.А., Рабулець О.Г., Сидорчук Н.М. Архітектура та компоненти програмного забезпечення системи українського національного лінгвістичного корпусу // MegaLing’2006. Горизонти прикладної лінгвістики та лінгвістичних технологій// Доповіді Міжнародної конференції. 2-27 вересня 2006, Україна, Крим, Партеніт / Ред. Широков В.А., Дікарєва С.С.; Український мовно-інформаційний фонд НАН України; Таврійський національний університет ім. В.І. Вернадського. – Сімферополь: Вид-во «ДиАйПи», 2006. – С. 26-27.

АНОТАЦІЯ

Сидорчук Н.М. Онлайнові лексикографічні системи. – Рукопис.

Дисертація у вигляді рукопису на здобуття наукового ступеня кандидата технічних наук за фахом 05.13.06 "Автоматизовані системи управління та прогресивні інформаційні технології". Національна бібліотека України імені В.І. Вернадського. – Київ, 2006.

У дисертації розвинуто теорію лексикографічних систем, орієнтованих на застосування Інтернет-технологій. На базі лексикографічної моделі даних, яка є основою для побудови лексикографічних систем, а також її узагальнень – лексикографічних середовищ та інтегрованих лексикографічних систем – викладено концептуальні моделі лінгвістичного корпусу, цифрової (електронної) бібліотеки з розширеними лінгвістичними функціями, а також інтегрованої лексикографічної системи «Словники України». Розглянуто онлайнові лексикографічні системи з погляду відкритих систем та обґрунтовано доцільність вибору сервіс-орієнтованої архітектури для онлайнових систем. Уперше розроблено принципи взаємодії архітектури OSI-7 та архітектури лексикографічних систем в ANSI/X3/SPARK інтерпретації. На базі викладених концептуальних та системотехнічних засад розроблено низку онлайнових Л-систем, серед яких: «Український лінгвістичний портал», інтегрована лексикографічна система «Словники України он-лайн», «Словник російської словозміни», «Цифровий архів документів Президії НАН України», мережевий варіант «Українського національного лінгвістичного корпусу» та ін.

Ключові слова: онлайнові лексикографічні системи, архітектура відкритих систем, архітектура лексикографічних систем, лексикографічні бази даних, інтегрована лексикографічна система «Словники України он-лайн».

АННОТАЦИЯ

Сидорчук Н.Н. Онлайновые лексикографические системы. – Рукопись.

Диссертация в виде рукописи на соискание ученой степени кандидата технических наук по специальности 05.13.06 "Автоматизированные системы управления и прогрессивные информационные технологии". Национальная библиотека Украины имени В.И. Вернадского. – Киев, 2006.

В диссертационном исследовании информационная теория лексикографических систем развита применительно к лексикографическим системам, ориентированным на функционирование в сетевых средах. На основании лексикографической модели данных, которая является основой для построения лексикографических систем, а также её обобщений – лексикографических сред и интегрированных лексикографических систем – изложены концептуальные модели лингвистического корпуса, цифровой (электронной) библиотеки с расширенными лингвистическими функциями, а также интегрированной лексикографической системы «Словари

Украины». Разработаны принципы взаимодействия архитектуры OSI-7 с архитектурой лексикографических систем в ANSI/X3/SPARK интерпретации. Отмеченное взаимодействие осуществляется на 5, 6, 7 уровне OSI-архитектуры с внутренним и внешним уровнем архитектуры Л-систем. Каждый из этих уровней обеспечивает сервис соседнему с ним верхнему уровню, получает сервис от смежного с ним нижнего уровня, осуществляет обмен блоками данных с целью осуществления определённых для этих уровней задач. Л-системы взаимодействуют с другими Л-системами, создавая тем самым лексикографическую среду, причём взаимодействие осуществляется между объектами одного и того же уровня. Рассмотрены онлайновые лексикографические системы с точки зрения открытых систем и обоснована целесообразность выбора сервис-ориентированной архитектуры для онлайнових систем. Изучены критерии качества для онлайновых лексикографических систем и проведено исследование их соответствия разработанным моделям и средствам. Среди них выделены: масштабируемость (онлайновая система должна предвидеть возможности колебаний нагрузки и реагировать на них); надежность (время простоя системы должно сводиться к минимуму); безопасность (онлайновая система должна быть защищена от несанкционированного доступа к ней); практичность (у разных пользователей должна быть возможность доступа к разному содержимому в разных формах); продуктивность (онлайновые системы должны иметь высокое быстродействие).

Исследовано четыре базовых стандарта, на основе которых осуществляется интеграция информационных систем с использованием веб-сервисов: расширяемый язык разметки, простой протокол доступа к объектам, язык описания веб-сервисов, универсальный метод описания, поиска и интеграции сервисов. На основании упомянутого исследования были разработаны системотехнические принципы реализации лексикографических систем в глобальной сетевой среде, а также выполнено проектирование и техническая реализация ряда онлайновых лексикографических систем. Среди них: «Украинский лингвистический портал», в составе которого разработана «Интегрированная лексикографическая система «Словари Украины он-лайн»; «Словарь русского словоизменения»; «Онлайновый вариант ведения научных журналов» (на примере журнала «Мовознавство»); «Цифровой архив документов Президиума НАН Украины», вариант электронной библиотеки с расширенными лингвистическими функциями и др.

Система «Словари Украины он-лайн» воссоздаёт в сетевой среде интегрированный лексикографический комплекс „Словари Украины” с присущими ему функциями словоизменения, синонимии, антонимии и фразеологии. Множество входов в систему не ограничивается реестровым рядом, но охватывает и правые части словарных статей. Каждое слово правой части является дополнительной точкой входа к соответствующей словарной единице. В пределах данного диссертационного исследования рассмотрены информационные потоки внутри системы,

организация интерфейса пользователя, проблемные аспекты реализации онлайнового программного комплекса.

На основании разработанной концептуальной модели цифровой библиотеки с расширенными лингвистическими функциями, системотехнических принципов построения онлайновых лексикографических систем создана специальная среда для сбора, хранения, моделирования и использования естественноязыковой информации в цифровом виде, которая стала подсистемой входа для Украинского национального лингвистического корпуса.

Ключевые слова: онлайновые лексикографические системы, архитектура открытых систем, архитектура лексикографических систем, лексикографические базы данных, интегрированная лексикографическая система «Словари Украины он-лайн».

ABSTRACT

N.M.Sydorchuk. Online lexicographical systems. – Manuscript.

The thesis is submitted to obtain an academic degree of candidate of technical sciences on specialty 05.13.06 “Computer-aided management systems and progressive informational technologies”. The Vernadsky National Library of Ukraine. - Kyiv, 2006.

In this thesis the theory of lexicographical systems oriented on the using Internet-technology was developed. The conceptual models of national linguistic corpus, digital library with extended linguistic functions, integrated lexicographical systems “The dictionaries of Ukraine” were described from the point of view of the lexicographical data model which forms the basis for lexicographical systems as well as for more general objects – lexicographical media and integrated lexicographical systems.

Online lexicographical systems were considered from the point of view of opened systems and expediency of choice of service-oriented structure for online systems was substantiated. For the first time principles of interaction of architecture OSI-7 and architecture of lexicographical systems in ANSI/X3/SPARK interpretation were developed. The series of online lexicographical systems like Ukrainian Linguistic Portal, Integrated Lexicographical System “The Dictionaries of Ukraine” online, Dictionary of Russian Inflection, Ukrainian National Linguistic Corpus, Digital achieve of documents for Presidium of Ukrainian National Academy of Sciences etc were developed on the basis of formed conceptual and systems engineering foundations.

Key words: online lexicographical systems, architecture of the opened systems, architecture of the lexicographical systems, lexicographical databases, Integrated Lexicographical System “The dictionaries of Ukraine on-line”.