Анатолій Широков, Олексій Костишин, Тамара Єрошенко

ПРОБЛЕМИ СТВОРЕННЯ ТА ВИКОРИСТАННЯ ЕЛЕКТРОННИХ БІБЛІОТЕК

Розглядається технологія формування електронних бібліотек. Вивчається проблема репрезентації гетерогенної інформації в електронній бібліотеці. Долсіджується питання створення спеціалізованих та проблемно-орієнтованих бібліотек. На прикладі електронної бібліотеки Українського мовно-інформаційного фонду ілюструється її лексикографічна спеціалізація.

1. ВСТУП.

Стрімкий розвиток комп'ютерної техніки, а саме персональних комп'ютерів лінії IBM PC, PS/2 та Apple Macintosh, високопродуктивних робочих станцій на основі RISC-процесорів (Sun SPARC, Dec Alfa та IBM/Apple Power PC, накопичувачів на оптичних дисках великого об'єму, розвиток швидких електронних мереж, які об'єднують значну кількість організацій на всій земній кулі) визначають новий підхід до бібліотек та бібліотечних систем.

Питаннями комп'ютеризації бібліотечної діяльності, створенням електронних каталогів, тобто інформаційно-пошукових комп'ютерних систем, що дозволяють за певними ознаками знаходити читачеві потрібну йому літературу та забезпечувати деякі допоміжні бібліотечні функції, в світі займаються вже давно.

Потреба в системах даного типу очевидна, оскільки навіть середня бібліотека має кілька десятків тисяч одиниць зберігання, є чимало бібліотек з мільйонними фондами. У найбільших бібліотеках світу число одиниць зберігання сягає десятків мільйонів.

Нові можливості щодо інформаційного пошуку створює й забезпечення доступу до електронних бібліотечних каталогів у режимах on-line та off-line по каналах комп'ютерного зв'язку. Такий сервіс надають не тільки визначні бібліотеки світу (Бібліотека Конгресу США, бібліотека Ватикану), а й середні за обсягами фондів університетські бібліотеки США, Німеччини, Англії, Франції та ін. Digital Equipment Corporation * для підвищення сервісу читачів запропонувала концепцію «бібліотеки без стін» (уперше було втілено в Тілбурзькому університеті (Нідерланди). У процесі реалізації цього проекту університетське містечко було зв'язане локальною комп'ютерною мережею з виходом на глобальну систему INTERNET, і всі користувачі вузівської бібліотеки одержали можливість комп'ютерного доступу до її каталогів та до бібліотек, підключених до глобальної системи. Аналогічні проекти реалізовано у ФРН (Геттінгенський університет), США (Бостонський та інші університети, Массачусетський технологічний інститут).

Але на сьогодні в світі існують не лише традиційні бібліотеки, оснащені електронними каталогами. Так, у межах Проекту Гутенберг (Gutenberg Project) * скарби світової літератури накопичуються у машинопрочитуваній формі (з 1974 р.). Доступ до цих матеріалів читач одержує, послуговуючися сервісом глобальної комп'ютерної мережі INTERNET, до якої нині під'єднано близько 3,5 млн. вузлів і десятки мільйонів користувачів з усього світу.

Велика комп'ютерна бібліотека текстів створюється в Інституті німецької мови (Institute fuer Deutsche Schprache) в Мангаймі (ФРН). Ще в 1964 р. засновники цього Інституту поставили завдання зробити німецьку мову як систему та соціальну структуру предметом лінгвістичних досліджень. У межах чотирьох проектів (їх веде згаданий Інститут за допомогою методів обчислювальної лінгвістики) провадиться не тільки накопичення творів класиків німецької літератури й газетних матеріалів у машинопрочитуваній формі, а й записується розмовна мова в цифровому вигляді.

Аналогічні ідеї розвиваються і в нас. У 1985 р. було започатковано науково-технічну програму «Машинний фонд російської мови та машинні фонди народів СРСР», виконання якої з розпадом Радянського Союзу припинилося.

2. РЕПРЕЗЕНТАЦІЯ ІНФОРМАЦІЇ В ЕЛЕКТРОННИХ БІБЛІОТЕКАХ.

З напрацюванням усе більшого об'єму інформації в машинопрочитуваній формі проблема розробки апаратно-програмного інструментарію, який давав би змогу виконувати більшість (або всі) бібліотечні функції, а також маніпуляції з текстами, які дозволяє сучасне програмне забезпечення, стає дедалі актуальнішою. Реалізація даної ідеї приводить до створення електронної бібліотеки, в якій у комп'ютерному вигляді зберігається не тільки вторинна інформація (бібліографічні описи),

*Одна з провідних корпорацій у справі комп'ютерно техніки та програмного забезпечення.

* Започаткований Іллінойським Бенедиктинським коледжем.

а й первинна (власне тексти книжок). З переходом до електронної бібліотеки фундаментальне поняття традиційних бібліотек - одиниця зберігання - підлягає значному узагальненню, зумовленому тим, що до звичайних видань та їх електронних аналогів дедалі частіше додаються аудіо- та відеокасети. Рисунки, графіки та інша невербальна інформація, яку включено до змісту видань, також потребує спеціальних засобів обробки. Отже, при проектуванні електронних бібліотек виникає проблема форматів зберігання та подання інформації у виданнях, де присутня гетерогенна інформація - текстова, графічна, аудіо, відео.

Особливої уваги потребує проблема багатомовності. Річ у тім, що в багатьох виданнях у змісті трапляються фрагменти, написані різними мовами. Навіть для читання таких текстів необхідно, щоб комп'ютер міг підтримувати мови, якими написані окремі фрагменти тексту. Коли ж справа доходить до змістовнішої роботи, скажімо, проведення автоматизованого аналізу тексту чи лексико-семантичних досліджень, то доводиться користуватися інструментарієм, який з самого початку орієнтовано на багатомовність. Розглянемо цю проблему докладніше.

Кожна сучасна природна мова L має скінченну множину символів, які зустрічаються при графічному (писемному) зображенні об'єктів даної мови. Називатимемо цю множину алфавітом над L і позначатимемо A(L). Такий алфавіт є дещо ширшим від звичайної абетки відповідної природної мови і має таку структуру:

$$A(L) = AL(L) \cup F(L) \cup P(L) \cup S(L) \cup \pi, \qquad (1)$$

де AL(L) є звичайний алфавіт природної мови L, записаний у стандартному порядку; F(L) - множина символів іншомовних алфавітів, які зустрічаються в розглядуваних текстах мови L і які не містяться в множині AL(L); P(L) - множина знаків пунктуації; S(L) - множина спеціальних символів (зокрема, арабських цифр, математичних, музичних символів тощо; π - пустий символ (пробіл).

Розглянемо узагальнення формули (1), в якому враховано можливість повної ідентифікації національних алфавітів. У цьому випадку член AL(L) $\cup$ F(L) у формулі (1) належить замінити на такий:

$$ALF(L) \equiv \bigcup_{i=0} AL(L_i), \qquad (2)$$

де $AL(L_i)$ - є звичайні алфавіти природних мов L_i, i=0,1, 2, ... , m, які зустрічаються у розглянутому тексті. В цьому випадку

$$L = \bigcup L_i, \qquad (3)$$

і загальна формула для знакової системи набуває такого вигляду:

$$A(L) = \bigcup_{i=0} AL(L_i) \cup P(L) \cup S(L) \cup \pi, \qquad (4)$$

У пропонованому підході однакові за графічним зображенням символи, що належать до різних алфавітів, мають різні системні номери при кодуванні і *a priori* розділяються. Отже, кожний символ поданого узагальненого алфавіту в системі однозначно задається значеннями двох параметрів, а саме: 1) своїм номером у кодовій таблиці та 2) номером самої таблиці, який ідентифікується з номером алфавіту $AL(L_i)$.

Проте в процесі автоматичної обробки багатомовних текстів з використанням оптичних методів постає проблема міжмовної омографії, яку не можна розв'язати на рівні кодових таблиць.

Схематично знакову систему типу (4) можна зобразити як набір з M примірників незалежних алфавітів, кожен з яких складається з певної кількості символів N_i, i=1, 2, ... , M. Тоді пара чисел (i, j), де i=1, 2, ... , M, j=1, 2, ... , N_i, є унікальним кодом довільного символу, який несе інформацію про належність цього символу до певної підмножини (підалфавіту) множин (4). У комп'ютері кожному коду (i, j) відповідає один і тільки один стан знакогенератора, який забезпечує відтворення графічного зображення відповідної літери на екрані монітора. Проте функція

$$G: (i, j) \rightarrow g_{ij}, \qquad (5)$$

яка визначає відповідність між кодом (i, j) та його графічним зображенням g_{ij}, хоча й є однозначною, але не є взаємнооднозначною. Отже, можуть виникати такі випадки, коли різним кодам відповідають тотожні графічні репрезентації. Іншими словами, на практиці може реалізовуватись (і дійсно реалізується) ситуація, коли

$$g_{ij} \equiv g_{kl} \quad \text{при } i \neq k \qquad (6)$$

Так, збігаються зображення літер «а», «о» в мовах, побудованих на латинській та кириличній графіках. Аналогічних прикладів можна навести чимало. При оптичному розпізнанні тексту програмне забезпечення OCR (Optical Character Recognition), що існує на даний момент, не може однозначно поставити у відповідність код літери в кодовій таблиці до її графічного образу. Щоб уникнути такої багатозначності, необхідна розробка спеціальних лінгвістичних і системних методів аналізу багатомовних текстів.

Реалізація знакової системи з описаними функціями не є простою. Про це свідчить досвід розвитку комп'ютерно-орієнтованих систем кодування символів, які вживаються на письмі. Від початку комп'ютери створювалися з орієнтацією на англійську мову, і першим стандартом на кодування письмових символів був Американський стандартний код для інформаційного обміну ASCII (American Standart Code for Information Interchange), що містив у собі лише великі та малі літери англійського алфавіту, цифри від 0 до 9, знаки пунктуації та деякі спеціальні символи. Цей код морально застарів, але й досі широко вживається, що пов'язано з його застосуванням у системах зв'язку. Розробка нового стандарту виявилася справою досить складною.

З поширенням комп'ютерів почали, по-перше, з'являтися проблеми в підтримці західноєвропейських мов, що не входять до англійської абетки. Тільки в 1977 р. Міжнародною Організацією Стандартів ISO було прийнято єдину 8-бітову кодову таблицю ISO 646 (вміщує символи західноєвропейських мов, символи псевдографіки та кілька математичних і фінансових знаків - усього 2^8 - 256). За основу стандарту ISO 646 було прийнято стандарт ASCII (виявився підмножиною ISO 646). Справді, перші 128 символів є символами ASCII. Не присутні в англійській графіці символи розміщено за номерами, більшими від 127.

Але проблеми країн, алфавіти яких не базуються на

латинській графіці, лишилися нерозв'язаними, зокрема країн з кириличним письмом, країн Близького Сходу, де використовується арабська писемність, і особливо - країн Далекого Сходу з ієрогліфічним письмом, що налічує десятки тисяч символів.

У 1983 р. консорціумом UNICODE було запропоновано двобайтову систему кодування і зберігання символів національних алфавітів, яка включає не тільки символи західно- та східноєвропейських алфавітів, а й символи івриту, арабської, грецької та східних мов. Розташування символів подається у кодовій таблиці.

Як бачимо, для підтримки подібної універсальності

КОДИ	СИМВОЛИ МОВ
від 0 до 8191 (000h - 1FFFh)	Абетки: англійська, європейські, фонетична, кирилиця, вірменська, іврит, арабська, ефіопська, бенгалі, девангарі, гур, гуджараті, орія, телугу, тамільська, каннада, малайська, сингальська, грузинська, тайська, бірманська, кхмерська, монгольська
від 8191 до 12287 (2000h 2FFFh)	Знаки пунктуації, математичні та технічні символи, орнаменти тощо
від 12288 до 16383 (3000h - 3FFFh)	Фонетичні символи китайської, корейської та японської мов
від 16384 до 59391 (4000h - E7FFh)	Китайські, корейські, японські ієрогліфи. Єдиний набір символів каліграфії хань.
від 59392 до 65023 (F800h - FDFFh)	Блок для приватного використання
від 65024 до 65535 (FE00h - FFFFh)	Блок забезпечення сумісності

розмір таблиці зберігання символів було збільшено з 256 до $2^{16} = 65536$ знакомісць. Зрозуміло, що це вимагає використання методів графічного зображення інформації, переходу до графічного режиму роботи дисплеїв і, отже, значного збільшення ресурсів комп'ютерів. Сучасні операційні середовища, такі, як Microsoft Windows, IBM OS/2, розробки груп фірми Digital Equipment і MIT - так званий інтерфейс користувача X/11 для UNIX-подібних систем, операційні системи комп'ютерів Apple Macintosh розраховані на графічну передачу інформації і дають змогу суміщати на екрані комп'ютера графічну й текстову інформації. Для цих операційних систем розроблено чимало прикладного програмного забезпечення: СУБД, текстові процесори, комп'ютерні видавничі системи, які користуються перевагами даного режиму.

Але в разі використання графічного режиму виникають певні ускладнення. По-перше, комп'ютери, побудовані на різних апаратних платформах, по-різному зберігають інформацію в машинному слові. По-друге, більшість згаданих програмних пакетів мають свої внутрішні формати зберігання даних і зв'язків між текстовими та графічними об'єктами. Внаслідок цього кожний такий програмний продукт доводиться наділяти конверторами і фільтрами для імпорту та експорту інформації та інших програмних систем. З метою певної стандартизації та спрощення розвитку й використання нових програмних продуктів фірмами Microsoft та Apple було розроблено платформонезалежний формат зберігання складних документів, прийнятий як стандарт де-факто, так званий формат RTF (Rich Text Format). Він фактично репрезентує формалізовану мову повного опису документа і використовує лише символи ASCII. Застосовуючи синтаксис даної мови, можна повністю описати досить складний документ, причому до цього опису ввійдуть формати його сторінок, типи застосовуваних шрифтів, їх розміри, посилання на ілюстративний матеріал, його поліграфічне виконання тощо. Отже, одержується платформо- та системно-незалежний документ, який можна обробляти на довільному комп'ютері за допомогою програмного забезпечення (реалізує інтерпретацію команд RTF). Підкреслимо, що останнім часом практично всі загальнопоширені системи обробки текстів забезпечуються конверторами з RTF до їх внутрішніх форматів незалежно від того, на якій апаратній платформі вони функціонують.

Ще одна проблема постає при узгодженні форматів збереження невербальних даних (графічної, аудіо- та відеоінформації). Не спиняючись окремо на аудіо- та відеоінформації, певного стандарту збереження яких ще навіть не запропоновано, розглянемо питання збереження статичної графічної інформації.

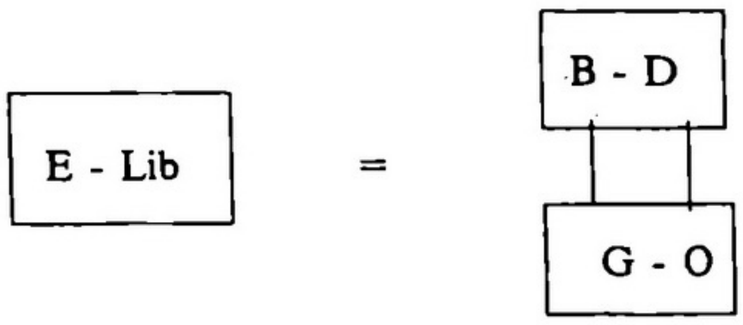
На сьогодні навіть для комп'ютерів IBM/PC існує значна кількість пакетів, кожен з яких по-своєму інтерпретує графічну інформацію, причому більшість з них мають бути настроєними на ту модель графічної плати (контролера графічної підсистеми), котра встановлена на комп'ютері, де цей пакет використовується. Через це перенесення графічної інформації навіть між комп'ютерами одного типу, але з різними графічними підсистемами, неможливе. Для запобігання цьому було розроблено кілька форматів репрезентації графічних даних, незалежних не лише від графічної підсистеми комп'ютерів однієї апаратної платформи, а й сумісних на комп'ютерах з різними апаратними платформами. До таких форматів слід віднести Grafic Interchange Format * або CompuServe GTF (широко використовується в системі глобальної інформації World Wide Web). Другим прикладом апаратно-незалежного формату збереження даних є Tagget Image File Format. Але платформна універсальність указаних форматів зумовлює їх дуже велику надлишковість. Так, тільки один рисунок або фотографія високої якості (роздільна здатність 1024x768 точок, 16 млн. кольорів) потребує для збереження буквально

* Запропонований консорціумом електронних мереж США CompuServe.

десятки мегабайт інформаційного простору. Отже, у виданні з 200-300 сторінками тексту і лише одним високоякісним рисунком сама текстова інформація займе менше одного відсотка загального об'єму. Об'єднаною групою експертів фотографів (Joint Photographer Expert Group) для суттєвого стиснення графічної інформації було запропоновано спеціалізований графічний формат збереження зображень високої якості. В цьому форматі інформація стискується в сотні і навіть тисячі разів без суттєвого погіршення якості. Завдяки цьому доля графічної інформації у гетерогенних об'єктах зменшується з 99 до 30-40 відсотків. Звичайно, декодування такої графічної інформації чисто програмними методами займає досить тривалий час (приблизно 10-15 хвилин на один рисунок), але вже існують суто апаратні компресори/декомпресори, які зменшують час декодування JPFG-зображень до секунд.

3. СТРУКТУРА ЕЛЕКТРОННОЇ БІБЛІОТЕКИ.

Як бачимо, комп'ютерна бібліотека поєднує в собі функції електронного каталога (ЕК) та засобів обробки гетерогенної інформації. Схематично це можна подати у такому вигляді:



(7)

У цій схемі B - D репрезентує блок бібліографічних описів ЕК, а G - O - блок узагальнених об'єктів зберігання, зображених у машинопрочитуваній формі (книжок, рисунків, аудіо-, відео-, графічної інформації, баз даних тощо). Розглянемо роботу зазначених блоків.

Вхідною підсистемою для електронної бібліотеки є блок бібліографічних описів або ЕК. Методи електронної каталогізації, як відзначалося у вступі, розвиваються вже досить давно. Вперше на промислову основу вони були поставлені в США. Так, ще в першій половині 60-х років у Бібліотеці Конгресу США (БК) було здійснено розробку програми машинопрочитуваної каталогізації MAPC (Machine-Readable Cataloguing), на основі якої утворилася найбільша міжнародна універсальна інформаційно-пошукова бібліографічна система. Формат MAPC дає змогу описувати не лише книжки, а й періодичні видання (журнали, газети тощо), географічні карти, ноти та ін. На основі формату USMARC надалі було розроблено бібліотечні формати для інших країн UKMARC (Великобританія), CANMARC (Канада) та ін. Різноманіття форматів зумовило розробку універсального формату для міжнародного обміну бібліотечними даними (UNIMARC). Формат ЕК Українського мовно-інформаційного фонду розроблений з урахуванням усіх вимог стандарту UNIMARC. Електронний каталог створено засобами пакету CDS/ISIS , версія 3.0. Вхідний формат складається з 54 полів з підполями, при глобальному коригуванні є можливість додати або відняти поле (підполе). Усі поля у базі змінної довжини. Експорт/імпорт здійснюється згідно стандарту ISO-2709. Введення даних у базу здійснюється засобами пакету CDS/ISIS , або за допомогою спеціального конвертора FONGORN. В ЕК використовується оболонка, яка забезпечує «дружній» інтерфейс користувача і допомагає йому здійснювати пошук за елементами бази даних та ключовими словами.

Технічно електронна бібліотека реалізована на комп'ютерах IBM PC, пов'язаних у локальну мережу, яка працює під управлінням сітьової операційної системи Novell NetWare V 3.12. На фізичному рівні зв'язок реалізовано коаксіальним кабелем. На транспортному рівні використовується протокол Ethernet 802.2. До файл-сервера підключено робочі станції користувачів, які працюють під операційною системою MS DOC у текстовому режимі, і основні пристрої накопичування та архівування інформації - накопичувачі на змінних оптичних дисках з перезаписом об'ємом 600 Мб та стример об'ємом 150 Мб.

Підсистеми зберігання узагальнених об'єктів працюють так. Електронні копії книжок після попередньої обробки (коректура, форматування тексту та архівація) надходять до електронного сховища. Ці видання бібліотека УМІФ одержує з різних видавництв і тому об'єкти електронної бібліотеки можуть сильно різнитися. По-перше, в багатьох видавництвах ще поширена видавнича система типу «КАСКАД» (реалізована на ЕОМ серії СМ-4). По-друге, завдяки розповсюдженню комп'ютерів лінії IBM PC чимало видавництв переходять на видавничі системи типу Xerox Ventura Publicher (як для DOC, так і для Windows), та Aldus Page Maker. У деяких видавництвах діють комп'ютери Apple Macintoch, на яких також використовуються системи типу Page Maker, а також поширена система QuarkXpress. При цьому об'єкти, які надходять до бібліотеки, мають різні формати. Найбільша кількість видань поки що надходить з системи «КАСКАД» на магнітних стрічках, записаних на пристрої СМ-5403 у форматі FLX в операційній системі ОС РВ. За допомогою спеціального пристрою і програмного забезпечення вони транспортуються і декодуються з формату FLX до звичайного формату ASCII-866, який можна читати довільним текстовим редактором. Складніше оперувати з текстами, підготовленими у Desktor Publishing System. Книга, передана в такій формі, має досить складний вигляд: це оригінал-макет, до якого включено інформацію про поліграфічний стиль видання, формати сторінок, графіки, рисунки тощо, причому вся ця інформація розміщена в різних файлах. Але на даний момент користувачеві для роботи можливо надати тільки чисті (так звані plain) тексти, тому ті машинопрочитувані видання, які надходять до бібліотеки у форматах класу сучасних видавничих систем, проходять первинну обробку (виділення текстової інформації). Одержані електронні копії книжок зберігаються і в оригінальному вигляді надходження. Така надмірність пов'язана з необхідністю збереження цілісності видання, яка може знадобитися для подальших лексико-семантичних та бібліографічних досліджень. Тобто, на сьогодні користувач має змогу працювати з текстовими частинами об'єктів збереження у звичайному текстовому режимі. Звісно, тут виникають ускладнення з багатомовними та гетерогенними об'єктами, про які вже згадувалося. Для розв'язання

цього зараз пристосовуються сучасні системні засоби до тематики електронних бібліотек з метою забезпечення підтримки графічних інтерфейсів.

До таких засобів слід, по-перше, віднести дуже поширену в світі надбудову над MS-DOC - MS-Windows. Для неї вже розроблено чимало різних інструментальних засобів - текстових процесорів, СУБД, Desktop Publishing Systems тощо, які дають змогу користувачеві оперувати гетерогенною інформацією. В цій системі ефективно реалізовано багатомовність за рахунок 2-байтової системи кодування. На жаль, зазначена система не підтримує стандарту UNICODE. Новіша реалізація системи MS-Windows™ - операційна система MS-Windows™ NT (підтримує стандарт UNICODE) перебуває в стадії випробувань. Однак розробка власних продуктів під системою MS-Windows™ є справою досить складною, оскільки фірма Microsoft, борючися з конкурентами, не документує значної кількості функцій ядра системи.

Альтернативою MS-Windows™ є 32-розрядна операційна система OS/2 фірми IBM. Вона набула популярності лише за останні 2-3 роки і для неї відсутній такий широкий спектр інструментальних засобів, як для MS-Windows™ . Крім цього, OS/2 не підтримує і стандарту UNICODE. Отже, попри велику популярність зазначених операційних середовищ, їх застосування для створення електронних бібліотек є проблематичним.

Перспективною для вказаних задач може стати операційна система UNIX , розроблена фірмою AT&T у середині 70-х років. Вона одразу проектувалась як багатозадачна і багатокористувацька система. Наприкінці 70-років робочими групами фірми Digital і MIT для системи UNIX було запропоновно первинний стандарт графічного інтерфейсу користувача X11. Основні його особливості такі:

- стандартизація органів управління програмними оболонками незалежно від розробника програмного забезпечення;
- вбудована підтримка локальних і глобальних мереж;
- реалізація і стандартизація міжпроцесорного обміну в термінах Клієнт- Сервер;
- вбудована підтримка багатомовності.

На відміну від MS-Windows™ та OS/2, система UNIX і специфікація протоколу X11 є відкритими. Відділення Каліфорнійського університету в Берклі розповсюджує систему UNIX та X11 на рівні вихідних текстів на мові програмування С.

4. ПРОБЛЕМНА ОРІЄНТАЦІЯ ЕЛЕКТРОННОЇ БІБЛІОТЕКИ.

Електронні бібліотеки, як і традиційні, можуть бути універсальними та спеціалізованими. Але якщо в традиційних спеціалізація визначається проблемною чи галузевою орієнтацією в доборі літератури, то спеціалізація електронної бібліотеки характеризується ще й специфічним набором програмних засобів, які дають змогу виконувати ті. чи інші операції над узагальненими об'єктами зберігання.

Так, електронна бібліотека УМІФ має лексикографічну спрямованість, що відбивається і в її структурі, і в спеціальній прикладній програмістиці, доступній її користувачеві. Вона орієнтована на дослідження української мови та укладання україномовних словників. Така спеціалізація спричинила створення спеціальної словникової підбібліотеки. В її електронному каталозі нині міститься понад 1500 бібліографічних описів словників. Для забезпечення більш релевантного пошуку в словниковій підсистемі довелося розширити (порівняно з усією бібліотекою) формат бібліографічного опису, не виходячи за межі стандарту ISO-2709.

Зокрема, було введено додаткові поля: шифр зберігання видання; інвентарний номер; ISBN. Крім викладеного, реалізується можливість підключення до бібліотеки деяких власне лексикографічних функцій.

Першою функцією (або підсистемою) є підсистема лексичної картотеки. Відомо, що однією з важливих напрямів лексикографічної роботи є добір ілюстративного матеріалу до вживання одиниць лексики в реальних мовних ситуаціях. Фіксованими джерелами такого матеріалу є опубліковані видання книжок, журналів, газет тощо - тобто системно це є узагальнені одиниці зберігання електронної бібліотеки, які в процесі їх створення вже пройшли належну філологічну апробацію. Ілюстративний мовний матеріал подається у вигляді відрізків тексту, що мають відповідати певним формальним критеріям.

Хоча ілюстративний мовний матеріал для лексикографа має службовий характер, достатньо повний набір прикладів вживання лексичних одиниць такий важливий для філологічної науки, що вже сама по собі проблема формування подібного набору набуває самодостатнього смислу і потребує розробки спеціальних методів його побудови та засобів маніпулювання відповідною інформацією.

Зазначена проблема, на перший погляд, не виглядає надто складною. Справді, видається, що досить переглянути певну кількість текстів (книжок), розбити їх на сегменти, які більш-менш однозначно розкривають зміст лексем, що містяться в цих сегментах, виписати або видрукувати ці сегменти на окремі картки і дати до кожної картки необхідну паспортизацію - тобто вказати заголовкове слово до кожної картки та літературне джерело, з якого її взято. Після цього розташувати одержані в такий спосіб картки за абетковим порядком заголовкових слів - і «інформаційна» система готова для експлуатації.

Але це - тільки з першого погляду. Насправді у ході формування та використання лексичної картотеки виникають труднощі, які практично неможливо подолати традиційними методами. Про що, зокрема, свідчить досвід формування лексичної картотеки української мови в Інституті мовознавства ім. О.О.Потебні НАН України. Створювана кількома поколіннями лексикографів, лексична картотека української мови (з 1991 р. вона функціонує в структурі Інституту української мови НАН України) нині містить понад 6 мільйонів карток. Обсяг інформації такий великий, а пошуковий апарат такою мірою нерозвинений, що ефективне використання картотеки унеможливилося. Контекст слів, що містяться в даній картці і не є заголовковими, залишається практично недосяжним, бо не існує пошукового апарату для їх локалізації в картотеці. Неможливим лишається і пошук за бібліографічними параметрами літературних джерел, з яких одержано лексичні картки. Це спричинює створення додаткових картотек та покажчиків до вже існуючої картотеки, що важко зробити через її великий обсяг. Отже, постає необхідність створення комп'ютерного варіанта лексичної картотеки, який був би вільним від описаних вад.

Нехай маємо деяку множину текстів (книжок):

$$T=\{ T_1, T_2, \dots , T_N\} \qquad (8)$$

Вважатимемо цю множину скінченною, але незамкненою, тобто такою, де можливе розширення шляхом додавання нових елементів. Позначимо через l_i довжину (в інформаційних одиницях, наприклад, у байтах) тексту книги T_i; через β_i позначимо бібліографічний опис цієї ж книги у прийнятому нами форматі.

Припустимо, що кожну книгу з множини T можна розбити на кілька частин тексту (сегментів), які в певному сенсі є самодостатніми для розкриття контексту вживаних у цих сегментах лексем. Отже, множина сегментів є такою:

$$\Sigma = \begin{matrix} \sigma^1_1, \sigma^2_1, \dots , \sigma^k_1 \\ \sigma^1_2, \sigma^2_2, \dots , \sigma^k_2 \\ \dots\dots\dots\dots \\ \sigma^1_N, \sigma^2_N, \dots, \sigma^k_N \end{matrix} \qquad (9)$$

де сегмент σ^j_i взято з тексту T_i. Вважатимемо, що алгоритм виділення сегментів з тексту є такий, що не допускає перетину між ними (у практиці, щоправда, зустрічатимуться випадки непустого перетину між деякими сегментами, але подібних випадків небагато і для наведених нижче міркувань вони неістотні). Тобто, загальна довжина всіх сегментів не перевищує повної довжини текстів:

$$\sum_{i=1}^{N} \sum_{r=1}^{k_i} |\sigma^r_i| \le \sum_{i=1} l_i \qquad (10)$$

В одному сегменті, як показує досвід, міститься до 50 слів. Оскільки середня довжина слів становить приблизно 5 літер, а на кодування однієї літери потрібен 1 байт інформації, то середня довжина одного сегмента тексту становить 2 кілобайти. Позаяк текстову інформацію можна зберігати у зкомпресованому вигляді, а програми компресії дають стиснення до 80%, то можна вважати, що довжина одного сегмента приблизно 400 байт. Отже, об'єм лексичної картотеки розміром у 6 мільйонів карток (такий об'єм лексичної картотеки Інституту української мови) мусить становити приблизно 2 - 2,5 Гбайт без урахування об'єму, який займає інформація пошукового апарату до системи. Зрозуміло, що бази даних таких розмірів вимагають розробки спеціальних методів їх формування і ведення, а також особливих вимог до методів зберігання даних, і до пошукового апарату. Принаймні, в процесі проектування системи слід розраховувати на базу даних не меншого об'єму інформації.

Яким же буде зазначений пошуковий апарат? Він має складатися з двох взаємопов'язаних частин.

Нагадаємо, кожному сегменту σ^r_i відповідає бібліографічний опис літературного джерела T_i, з якого його взято. Цей опис є однозначною функцією нижнього індекса і, який є ідентифікатором тексту (книги). Отже, одна частина пошукового апарату являє собою бібліографічний опис, сконцентрований у системі пошукових індексів $\{\beta_i\} = \beta_\Sigma$, коли йдеться саме про ту множину текстів, з яких виділено систему сегментів Σ.

Друга й складніша частина пошукового апарату - це словникова система, параметризована так, що більшість її інформаційних змінних є пошуковими. Розглянемо докладніше функції та структуру цієї системи.

Відповідно до сучасної методології створення лексичних картотек у процесі їх формування повинен підтримуватися принцип суцільного розписування тексту. Тобто, довільна лексема у виділеному в книзі сегменті функціонально є пошуковою. Реалізація згаданого принципу передбачає такі операції:

1. Редукція всіх словоформ, присутніх у наявному наборі сегментів $\{\sigma^r_i\}$, до вихідних, початкових форм (лематизація);

2. Побудова словника V (Σ) - алфавітного списку всіх лексичних одиниць, присутніх у наборі $\Sigma = \{\sigma^r_i\}$, у початковій (канонічній) формі;

3. Індексація словника за номерами-ідентифікаторами карток σ^r_i: зіставлення кожного слова словника номерів усіх карток, у яких присутні всі елементи парадигми даного слова, а також реалізація пошукових функцій. Тобто, нехай x_i - якесь слово в початковій формі - елемент словника, описаного у п.2; позначимо через $[x_i]$ його повну парадигму. Тоді базова пошукова операція буде такою: знайти всі сегменти σ^r_i, в яких зустрічається словоформа s_{i_r} що належить до $[x_i]$.

4. Розширення словника з надходженням нових сегментів до лексичної картотеки (фіксація нових слів, контроль за їх надходженням та ведення статистики).

Позначимо для простоти через V (Σ) словник, який містить слова, наявні в системі Σ, разом з його пошуковим апаратом та іншими функціями, описаними у пп. 1-4. Тоді лексична картотека являє собою таку сімку:

$$ЛК = \{ \Sigma = \{\sigma^r_i\}; A(T); V(\Sigma); \Lambda; \Pi; \{\beta(\Sigma)\}; S \}, \qquad (11)$$

де $\Sigma \{\sigma^r_i\}$ - множина сегментів; A(T) - множина алгоритмів сегментації текстів (або, більш загально, алгоритмів маніпуляцій над текстами); V (Σ) - словник з відповідним пошуковим апаратом; Λ - множина алгоритмів лематизації; Π - множина алгоритмів парадигматизації;{β (Σ)} - бібліографічний пошуковий апарат; S - множина функцій системного сервісу.

Наступна функція стосується технології створення словників і базується на розробленій в УМІФ лексикографічній моделі словникових систем, яка спирається на багатоаспектне зображення знакової природи лексичних одиниць як найбільш компактних та інформативних у будь-якій природній мові. Вихідним моментом при побудові лексикографічної моделі, її своєрідним нульовим наближенням є врахування в слові дихотомічної структури в багатовимірному, статичному відбитті співвідношення його форми та змісту.

Статичне зображення дихотомічної природи слова зумовлюється насамперед фіксованою формою подання змісту в словнику: будь-який словник здатен фіксувати не всю, а лише обмежену кількість реєстрових одиниць, не всю історію розвитку конкретної мови, а тільки більш-менш тривалий відтинок часу, не всі, а лише деякі сфери її вживання. Досі ще не існує словників, які б віддзеркалювали сам процес мовотворення, а не тільки його результати.

Багатовимірність співвідношення форми та змісту слова як мовної одиниці в паперових словниках забезпечується урахуванням фонетичних,

морфематичних, граматичних, семантичних та стилістичних особливостей слова. В розроблюваній моделі вони описуються такими структурами:

$$\begin{aligned} &D_{Ph}[S(L)] \text{ - фонетичний опис;} \\ &D_{Morph}[S(L)] \text{ - морфематичний опис;} \\ &D_{Gr}[S(L)] \text{ - граматичний опис;} \\ &D_{Sem}[S(L)] \text{ - семантичний опис;} \\ &D_{St}[S(L)] \text{ - стилістичний опис,} \end{aligned} \qquad (12)$$

де через S(L) позначено множину слів конкретної природної мови L, описаних опублікованими досі словниками і включених у розроблювану лексикографічну модель. Зазначена множина вважається відкритою, але скінченною. Під описом (description) розуміється тут відображення предметної області (в даному випадку предметною областю є «СЛОВО») у деяку множину моделей даних чи знань.

Отже, дихотомічність структури слова в розроблюваній лексикографічній інформаційній моделі зображена за зразком більшості опублікованих словників, де в реєстрі повнозначні частини мови подані вихідними (канонічними) словоформами, а в моделі становлять підмножину $S_0(L) \subset S(L)$. З наведеного випливає можливість зображення кожної структури системи (12), а також довільного їх об'єднання та перетину у вигляді декомпозиції:

$$D[S(L)] = \{S_0(L); D_F[S(L)], D_C[S(L)]\}, \qquad (13)$$

де через $D_F[S(L)]$ позначено формальну частину опису множини $S_0(L)$, а через $D_C[S(L)]$ - її змістову частину.

Узагальненість дефініцій $D_F[S(L)]$ та $D_C[S(L)]$ дає змогу досить оперативно користуватися системою (13) при лексикографічній кваліфікації слова і як одиниці тексту, і як одиниці словника. Наприклад, фонетичний опис в одних випадках може виступати у функції $D_F[S(L)]$, а граматичний - виконувати роль змістовної частини - $D_C[S(L)]$; в інших випадках функцію $D_F[S(L)]$ виконуватиме саме граматичний опис, тоді як у ролі $D_C[S(L)]$ виступатиме семантичний опис тощо.

Зазначений факт для лексикографічної моделі є досить фундаментальним, бо завдяки йому порівняно легко й оперативно здійснюється настроювання моделі на укладання словників різних типів чи використовування їх як внутрішніх елементів інформаційних систем. По-різному укладаються та використовуються в практичній діяльності, наприклад, тлумачний словник з його одномовними, але багатокомпонентними реєстровою та тлумачною частинами, термінологічні й загальномовні перекладні словники, в яких зіставляються різномовні реєстрова і перекладна частини словникових статей, орфоепічний та орфографічний словники, в яких граматичний опис реєстрових слів виділено в окрему структурну частину словникової статті, а реєстрова одиниця, як правило, однокомпонентна і т. ін. Відзначені особливості варіювання величин $D_C[S(L)]$ та $D_C[S(L)]$ являють собою ті детермінанти, через які в створюваній автоматизованій системі визначаються тип кожного конкретного словника та його лексикографічна й дексикологічна спрямованість.

Дихотомічність структури слова як мовного знака зображується в моделі за допомогою виділення у словникових статтях реєстрової та інтерпретаційної частин. У літературі реєстрова частина називається лівою, а інтерпретаційна - правою. Звідси загальна формула зображення будь-якого словника в лексикографічній моделі матиме такий вигляд:

$$V(L) = \{\Lambda(L); P(L), H\} \qquad (14)$$

В останній формулі через V(L) позначено словник як множину словникових статей; $\Lambda(L)$ є множиною лівих частин словникових статей словника V(L); P(L), відповідно, - множиною правих частин цього ж словника; H - відображення множини $\Lambda(L)$ на P(L):

$$H: \Lambda(L) \rightarrow P(L) \qquad (15)$$

Отже, H виступає функцією, яка ставить у відповідність лівій частині словникової статті, її праву частину і забезпечує дихотомічну цілісність побудови тексту відповідної словникової статті.

Зауважимо, що терміни «ліва» й «права» частини є до певної міри умовними, оскільки межа між фрагментами структури словникової статті які позначаються цими термінами, не завжди однолінійна. Лексикографічне розмежування лівої і правої частин, отже, стосується не так формально позиційного їх розташування в словниковій статті, як зображення функціонального протиставлення форми та змісту в слові. В паперових словниках у їх безпосередньому друкарському виконанні часто трапляються випадки перемішування окремих елементів структури $\Lambda(L)$ і P(L). Ще ясніша умовність цих назв у разі користування комп'ютерними словниковими системами.

У структурі паперових словників, як і в лексикографічній моделі, словникова стаття починається відповідним реєстровим словом x, що разом з тим, вважається її ідентифікатором. Отже, формулу (14) словника можна деталізувати в такий спосіб:

$$\begin{aligned} &V(L) = \bigcup_{\chi \in S_0(L)} V(\chi); \quad \Lambda(L) = \bigcup_{\chi \in S_0(L)} \Lambda(\chi); \\ &P(L) = \bigcup_{\chi \in S_0(L)} P(\chi) \end{aligned} \qquad (16)$$

де V(L) - словникова стаття, яка очолюється реєстровим словом χ, а $\Lambda(X)$ і $P(X)$ - відповідно ліва і права частини цієї словникової статті. Таким чином:

$$H(\Lambda(\chi)) = P(\chi).$$

На множині V(L) визначається частковий порядок, індукований «лексикографічним» упорядкуванням множини S(L).

Для словника V(L) може існувати автоморфізм, тобто відображення:

$$A: V(L) \rightarrow V(L), \qquad (17)$$

яке констатує наявність відсилочних словникових статей типу: X див. Y. Зазначений автоморфізм визначає таке зображення словникових статей $V(X) \rightarrow V(Y)$. Його ідентифікатором є, як правило, якесь відсилочне слово (символ, абревіатура; у наведеному прикладі - «див.»), що зіставляє словниковій статті V(C) її відповідник V(Y). Але будова автоморфізму A може бути складнішою за наведену в нашому

прикладі По-перше, довжина низки відсилань може бути більшою за одиницю, тобто мати ланцюгово розгортальний вигляд:

$$V(X) \rightarrow V(X') \rightarrow \ldots \rightarrow V(X'')$$

Крім того, зображення $V(X) \rightarrow V(Y)$ може репрезентувати цілий пучок відсилань. Це реалізується, наприклад, коли стаття V(x) має таку будову:

$$X(X', X'', \ldots) \text{ див. } V(V', V'', \ldots)$$

Тут в одній словниковій статті V(L) визначено пучок зображень:

$$V(X) \rightarrow V(Y);\ V(X') \rightarrow V(Y');\ V(X'') \rightarrow V(Y''); \ldots$$

Отже, зображення H та A породжують макроструктуру словника V(L). Крім цього, кожен словник має ще й свою внутрішню мікроструктуру, яка відбиває у неявному вигляді семантику предметної області, що є об'єктом конкретного словника. Вказана мікроструктура стосується будови об'єктів Λ(x) і P(x).

Лексикографічне моделювання словникових статей показує, що розвинений формалізм спроможний повністю відтворити всі деталі структури будь-якого словника. Це дає змогу розробки технологічної системи для укладання словників різних типів завдяки автоматизації таких функцій: генерація узагальненої абетки словника; породження структури словникових статей конкретного словника і формування відповідної бази даних; граматична та лексико-семантична ідентифікація об'єктів словника, включаючи функції лематизації та автоматичної побудови парадигми для повнозначних відмінюваних частин мови, проведення лексикологічних досліджень; злиття словників та одержання підсловника з даного словника; виконання коректорських і редакторських робіт; одержання оригінал-макета готового словника, а також цілий ряд сервісних інформаційно-пошукових функцій.

1. Unimarc: Universal MARC format. 2 nd ed. - London, 1980. - XIV. - 131 p.
2. ГОСТ 7.1-84. Библиографическое описание документа. Общие требования и правила составления. - М.: Изд-во стандартов, 1984. - 77 с.
3. Формат библиографических данных. Состав и наполнение полей. Всес. книжная палата. - М., 1990. - 33 с. (машинопись).
4. Библиотека и библиотечное дело США: комплексный подход / Под ред. В.В.Попова.

Ювілеї бібліотек

Ельга Татарчук

75 РОКІВ БІБЛІОТЕЦІ УКРАЇНСЬКОГО ДЕРЖАВНОГО ПЕДАГОГІЧНОГО УНІВЕРСИТЕТУ ІМ. М.ДРАГОМАНОВА

За 75 років існування наш вуз випустив понад 65 тисяч учителів і працівників системи дошкільного виховання. Сьогодні на 11 його факультетах готуються педагогічні кадри з 24 спеціальностей.

Бібліотеку було засновано одночасно з Київським інститутом народної освіти ім.М.Драгоманова (таку назву мав інститут у 20-ті роки). До 1935 р. її книжковий фонд становив 100 тис. примірників. Перед Великою Вітчизняною війною він зріс до 170 тис. одлб., що, в основному, уціліли. Швидко збільшувався масив літератури у повоєнний час. У 1953 р. він становив 250 тис. книг, а кількість читацьких місць сягнула 120 (порівняємо: у 30-ті роки - 20).

З будівництвом нових навчальних корпусів (у 70-х роках) у них були відкриті факультетські бібліотеки.

Сьогодні бібліотека Українського державного педагогічного університету ім.М.Драгоманова найбільша серед бібліотек вищих педагогічних установ країни і є для них базовою. Її фонд налічує 1 млн. 380 тис. прим. книг, журналів, дисертацій, мікрофільмів. На 6 абонементах і в 12 читальних залах обслуговується понад 16 тис. студентів, аспірантів, викладачів, учителів Києва й області. Щоденно - понад 2 300 читачів.

Перехід вузу на багатоступеневу освіту, поновлення її змісту, по-сучасному організована самостійна робота студентів перетворюють бібліотеку в один з найактивніших учасників навчального процесу. Найбільше відвідувачів - у спеціалізованих читальних залах. Так, працівники читального залу суспільно-гуманітарного профілю допомагають добирати матеріали до семінарських занять, здійснюють огляди літератури з лекційних тем, влаштовують виставки нових книг і журналів. Постійно поповнюється і дедалі більше привертає увагу читачів зал літератури з українознавства. Книжкові фонди бібліотеки розкриті за змістом у системі каталогів (генеральний алфавітний, систематичний, алфавітні та систематичні каталоги дисертацій, авторефератів, іноземної літератури, періодичних видань). Ретельно розроблені традиційні каталоги доповнюються електронним каталогом на нові надходження (з 1992 р.), для чого застосовується формат MARC. Співробітництво вузу з відомою фірмою Digital Edulpment Corporation дасть змогу комп'ютеризувати бібліотеку і ввійти до «Віртуальної бібліотеки Києва» з перспективою інтеграції у всесвітню інформаційну мережу INTERNET.

З 1992 р. бібліотека розпочала створення національної педагогічної бібліографії. Видано 4 випуски покажчика (тираж щорічника - 500 примірників) «Українська педагогічна бібліографія» (література за 1990 - 1993 рр.), бібліогрфічні покажчики друкованих праць співробітників університету за 1944 - 1989 рр., «Професійна орієнтація учнів», «Образ учителя в художній літературі» та ін.

Культурно-виховні заходи, у тому числі книжково-ілюстровані виставки (понад 120 щороку), які проводить бібліотека, спрямовані на популяризацію літератури, яка з сучасних позицій висвітлює питання історичного розвитку України, її культури. У читацьких конференціях, літературних вечорах, усних журналах беруть участь майстри художнього слова, відомі київські актори.

Здійснюючи інформаційне забезпечення навчальної, наукової та виховної діяльності педуніверситету, бібліотека, яку часто називають головною лабораторією вузу, надає неоціненну допомогу у підготовці нової інтелектуальної еліти.