УДК 658.012.01156

І. М. Білятинська
Управління комп'ютеризованих систем апарату Верховної Ради України

ІНТЕГРОВАНІ ЛЕКСИКОГРАФІЧНІ СИСТЕМИ ТА МЕТОДИ ЇХ ПОБУДОВИ

Розглянуто основні аспекти побудови інтегрованих лексикографічних систем, описано особливості інтеграції на концептуальному, внутрішньому, зовнішньому рівнях лексикографічних систем та запропоновано послідовність кроків, які можна використовувати для реалізації цього процесу.

Ключові слова: лексикографічна система, інтеграція, інформаційна система, типи інтеграції, інтеграція електронних словників, інтегрована лексикографічна система.

In this paper, we examine the main aspects of the integrated lexicographical systems. We describe the features of the integration on conceptual, internal and external levels of lexicographical systems and propose sequence of steps that can be used to integrate lexicographical systems.

Key words: lexicographical system integration, information systems, types of integration, integration of electronic dictionaries, lexicographical integrated system.

Вступ

Питання інтеграції інформаційних систем, незважаючи на його доволі глибоке опрацювання, залишається і досі актуальним. Те саме стосується й інтеграції електронних лінгвістичних засобів – лексикографічних систем (Л-систем) та електронних словників як елементів лінгвістичних технологій, оскільки темпи впровадження та, власне, модус цих технологій у сучасному мережевому середовищі набувають все більших масштабів та обширів застосування. Зазначимо, що питання інтеграції Л-систем свого часу докладно розробили В. А. Широков та О. Г. Рабулець, воно основане на створеній В. А. Широковим теорії лексикографічних систем та (спільно з О. Г. Рабульцем) теорії лексикографічних середовищ [2, 4]. Підсумок цієї роботи викладено в дисертації О. Г. Рабульця "Інтегровані лексикографічні системи" [2]. Також зазначимо, що згадані теоретичні підходи практично втілено в низці конкретних цифрових лексикографічних праць, зокрема, в унікальному мережевому мовно-інформаційному об'єкті "Інтегрована лексикографічна система "Словники України" [3], у якому для української мови зінтегровано функції побудови повної словозмінної парадигми, фонематичної транскрипції, синонімії, антонімії та фразеології, причому на колосальному лексичному обсязі понад 250 тис. одиниць; дещо спрощена версія цієї системи доступна в Інтернеті за посиланням http://lcorp.ulif.org.ua/dictua/.

Проте, незважаючи на серйозні досягнення в цій галузі, значна частина важливих теоретичних і практичних аспектів інтеграції мовно-інформаційних об'єктів поки що залишається не розглянутою. Та й узагалі, проблема інтеграції, на нашу думку, належить до вічних проблем системотехніки, які час від часу необхідно розглядати з оновлених позицій та з урахуванням набутого за певний період досвіду. Зазначені аспекти в цій статті розглянемо на прикладі інтеграції трьох дуже складних лексикографічних систем, які створено в Українському мовно-інформаційному фонді НАН України у вигляді цифрових інтерпретацій тлумачного Словника української мови, Граматичного словника української мови та Етимологічного словника української мови. Програмні реалізації цих Л-систем виконано із використання тих самих підходів до розроблення: клієнтська частина реалізована в середовищі розробки Microsoft Visual 2015

Community з використанням базового фреймворку .Net framework 4.6, веб-сервіси реалізовано з використанням ASP.NET WebApi 2, користувацький інтерфейс – Windows Presentation Foundation (WPF), розгортання системи – ClickOnce. Лексикографічні бази даних всіх трьох об'єктів використовують СУБД Microsoft SQL Server 2014 Express.

Виклад основного матеріалу

Інтегровані лексикографічні системи та методи їх побудови

Необхідно, однак, зауважити, що питання побудови цифрових реалізацій комплексних Л-систем, у яких поєднуються різнорідні мовні феномени та ще й на повному обсязі їхніх функцій і лінгвістичних параметрів, далеко виходить за межі суто програмної проблематики і потребує застосування моделей та засобів, набагато загальніших, ніж власне програмні. Такий загальний план розгляду, дослідження та проектування, на нашу думку, сконцентровано в понятті *архітектури інформаційної системи*. Услід за нашими попередниками, використовуємо трирівневу інформаційну архітектуру ANSI/X3/SPARK [1], проте в дещо докладнішій, ніж в оригіналі, інтерпретації, узгодженій з визначенням поняття лексикографічної системи.

Фундаментальне поняття Л-системи описує доволі загальний тип формалізованих конструкцій, поряд з моделями даних, формальними граматиками тощо. В науці та техніці прикладами конкретних реалізацій лексикографічних систем є різноманітні інформаційні системи, бази даних, словники в традиційній чи електронній формі, енциклопедії тощо. Під час проектування інформаційно-лінгвістичних об'єктів використання Л-систем дає змогу врахувати специфіку мовних явищ, відобразити в найповнішому обсязі форму та зміст одиниць мови.

Під *лексикографічною системою (Л-система)* мають на увазі інформаційний об'єкт, який поєднує в собі риси моделі даних, моделі знань, логіко-лінгвістичного числення певного типу та являє собою спеціальне інформаційне середовище, в якому розвивається лексикографічний ефект (або певна їх сукупність) [4, с. 109]. Поняття лексикографічного ефекту ґрунтується на твердженні, що в процесі функціонування систем будь-якого типу в її структурі можна виокремити підсистему порівняно сталих дискретних сутностей, які відіграють роль її елементарних інформаційних одиниць так, що систему можна описати як сукупність комбінацій цих елементарних інформаційних одиниць та визначає процес, унаслідок якого згадана сукупність є сталою та локалізується у відповідних областях системних параметрів. Універсальність цього поняття дає змогу розглядати дослідження будь-яких предметних галузей як вивчення лексикографічних ефектів, характерних для явищ, об'єктів чи процесів, що стали предметом цих досліджень.

Лексикографічна система являє собою певну лексикографічну модель даних разом з конкретною її реалізацією. Загальний випадок лексикографічної моделі даних описується вісімкою об'єктів:

$$\{I^Q(D),\ S,\ V(I^Q(D)),\ \beta,\ \sigma[\beta],\ RR{\downarrow}[V(I^Q(D))]\},$$

де $I^Q(D)$ – клас елементарних інформаційних одиниць (ЕОМ), S – суб'єкт сприйняття, $V(I^Q(D))$ – множина опису ЕОМ, β – структури, характерні для описуваного явища, $\sigma[\beta]$ – макроструктури $V(I^Q(D))$, $RR{\downarrow}[V(I^Q(D))]$ – процес рекурсивної редукції Л-системи. Реалізація лексикографічної системи найчастіше являє собою інформаційну систему, архітектура якої позначається символом Σ.

Отже, зміст загального визначення лексикографічної системи відображається за допомогою конструкції:

$$\{I^Q(D),\ S,\ V(I^Q(D)),\ \beta,\ \sigma[\beta],\ RR{\downarrow}[V(I^Q(D))],\ \Sigma\}.$$

В останній формулі символом Σ позначено архітектуру Л-системи, розглядуваної як інформаційна система певного типу. Як зазначено вище, ми застосовуємо трирівневу архітектуру ANSI/X3/SPARK, що складається з трьох рівнів абстракції даних: концептуального, внутрішнього та зовнішнього:

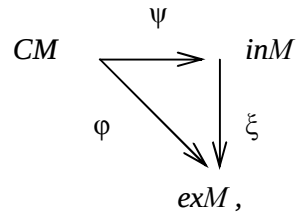
$$\text{ARCH\_LS} = \{\text{CM, EXM, INM; } \Phi, \Psi, \Xi\}.$$

В останній формулі символом CM позначено концептуальну модель Л-системи, EXM = {exM} – множина зовнішніх, а INM = {inM} – множина внутрішніх моделей, які

відповідають цій концептуальній моделі СМ. Через Φ = {φ}, Ψ = {ψ}, Ξ = {ξ} позначено множину відображень СМ у EXM, INM, EXM відповідно:

$$\varphi : CM \rightarrow exM, \text{ де } exM \in EXM;$$
$$\psi : CM \rightarrow inM, \text{ де } inM \in INM;$$
$$\xi (inM) = exM, \text{ де } exM \in EXM.$$

Це явище ілюструє діаграма:



яка є комутативною, тобто: ξ ψ = φ. Вимога комутативності діаграми істотна, оскільки забезпечує узгодженість між різними рівнями представлення даних. У цій архітектурі ми зупиняємося на такій інтерпретації рівнів представлення даних:

Концептуальний рівень являє собою однозначний, скінченний, вичерпний, несуперечливий опис об'єктів лексикографічної системи та зв'язків між ними.

Внутрішній рівень – це загальний опис структур даних, що характерні для лексикографічної системи, а також правила та засоби маніпулювання ними. Цей рівень охоплює фізичне подання даних, їхню модель та конкретну систему управління базами даних.

Зовнішній рівень – сукупність програмного забезпечення, за допомогою якого реалізується доступ кінцевих користувачів до даних внутрішнього рівня лексикографічної системи. Зовнішній рівень містить API (*Application programming interface*, прикладний програмний інтерфейс), і користувацький інтерфейс.

Під час побудови тієї чи іншої лексикографічної системи з усієї складної ієрархії лексикографічних ефектів, які природно існують, розглядаються лише ефекти певного типу, що є предметом дослідження, інші ж можуть стати основою для побудови інших систем. Такий підхід виправданий і дає змогу зосередити увагу на вивченні окремих властивостей мови, тим самим обмежившись оптимальним рівнем складності системи.

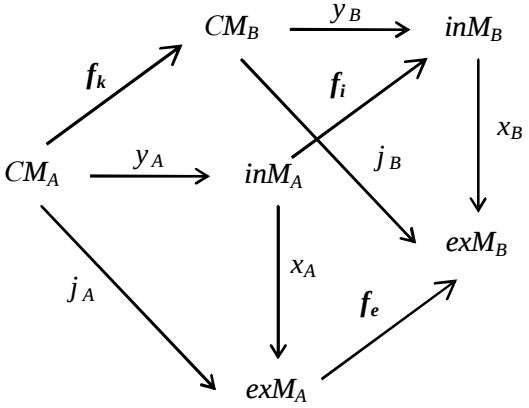
У ході вирішення численних лінгвістичних завдань з'являється потреба у можливості системного дослідження різних типів лексикографічних ефектів, а отже, необхідність інтеграції лексикографічних систем, здебільшого неоднорідними за всіма рівнями архітектури. Для виконання цього завдання виникає потреба у побудові лексикографічного середовища, що являє собою мовно-інформаційне середовище із визначеними засобами та конструктивами, необхідними для здійснення процесів інтегрування різних Л-систем та фіксації їх результатів у вигляді інтегрованих Л-систем.

Завдання інтеграції Л-систем, неоднорідних за всіма рівнями архітектури, потребують розроблення методів інтеграції концептуальних моделей, способів подання даних та операційно-програмних платформ, а також узгодження зовнішніх представлень відповідних концептуальних схем та їх внутрішніх репрезентацій [4, с. 128]. Результатом цієї діяльності є інтегрована лексикографічна система, що являє собою об'єднання кількох програмних засобів, які працюють разом, що в результаті інтеграції надають своєму користувачеві можливості, не притаманні кожній з них поодинці.

Лексикографічні системи, які потребують інтеграції, розглядатимемо як об'єкти лексикографічного середовища. Найзагальніший випадок процесу інтеграції двох об'єктів А та В можна проілюструвати діаграмою, наведеною на рисунку, де CM_A, CM_B – концептуальні моделі А та В, inM_A inM_B множини внутрішніх, а exM_A exM_B – множини зовнішніх моделей, які відповідають концептуальним моделям А та В відповідно; f_c, f_i, f_e компоненти вектора $f: A \to B$, $f \hat{I} Hom_{ML}(A, B)$, де $Hom_{ML}(A, B)$ – множина морфізмів А у В.

Інтеграція може здійснюватися на всіх трьох рівнях одночасно або на певному з них, залежно від особливостей компонентів інтеграції та цілей цього процесу.

Інтеграція на концептуальному рівні дає змогу абстрагуватися від структур даних та конкретних програмних реалізацій лексикографічних систем, зосередивши увагу на описі об'єктів лексикографічних систем у термінах єдиної моделі, виявленні явищ, що стали об'єктами лексикографування одночасно в кількох компонентах інтеграції, визначення того, як відомості, подані в одній лексикографічній системі, можуть доповнити та розширити відомості, представлені в інших.



Інтеграції двох об'єктів A та B

Інтеграція на внутрішньому рівні лексикографічної системи найчастіше являє собою інтеграцію на рівні даних. Така інтеграція потребує подання структур даних, які використовують лексикографічні системи, що інтегруються, у термінах єдиної моделі. Кожна з них може мати власні реалізації концептуального та зовнішнього рівнів презентації. Складність такої інтеграції залежатиме від різнорідності даних, що потребують інтеграції.

Інтеграція на зовнішньому рівні – це завжди інтеграція конкретних програмних реалізацій лексикографічних систем.

Оскільки лексикографічні системи можуть мати різноманітну структуру, в їх основу можуть бути покладені різні лексикографічні ефекти, вони можуть відрізнятися рівнями складності, підходами до своїх програмних реалізацій, тому не можна сформулювати універсальне правило, як саме потрібно їх інтегрувати. Проте, на нашу думку, процес інтеграції лексикографічних систем може бути здійснений у результаті виконання таких кроків.

1. Побудова концептуальних моделей лексикографічних систем (компонентів), що стали об'єктом інтеграції.

2. Аналіз отриманих концептуальних моделей.

3. Виявлення явищ, одночасно описаних у двох та більше компонентах. Визначення, як ці описи перетинаються та як доповнюються. На основі цих відомостей визначити, які нові властивості описаних явищ можна отримати внаслідок інтеграції.

4. Побудова нової концептуальної моделі на основі отриманих відомостей, що являє собою інтеграцію концептуальних моделей, побудованих на кроці 1.

5. Оцінити складність отриманої концептуальної моделі та визначити, як реалізовувати подальшу інтеграцію лексикографічної системи: на внутрішньому, зовнішньому рівнях чи одночасно на обох з них. У деяких випадках, коли жоден з цих варіантів неможливо застосувати, доцільно на основі отриманої концептуальної моделі реалізувати нову лексикографічну систему, із власними внутрішнім та зовнішнім рівнями представлення.

6. У разі доцільності інтеграції на внутрішньому рівні спочатку виявляють елементи структури лексикографічних моделей, які мають однакове семантичне значення у двох та більше з них (семантично тотожних атрибутів) й відповідних їм областей визначення та елементів структури, що мають різне семантичне значення для різних лексикографічних моделей. З метою зняття омонімії вводяться нові атрибути та механізми заміни ними потрібних елементів структури.

7. На основі аналізу, здійсненого на попередніх кроках, вибрати тип програмної реалізації інтегрованих лексикографічних систем. Підходи та методи програмної реалізації інтегрованих лексикографічних систем, з огляду на складність та обсяг матеріалу, розглядатимемо в наступних роботах.

Висновки

Процес інтеграції лексикографічних систем ускладнюється необхідністю об'єднання складних об'єктів, які можна реалізувати за допомогою різного програмного забезпечення, використовувати різноманітні структури даних, а головне, в їх основу можуть бути покладені різні лексикографічні ефекти. Тому під час вирішення цього завдання необхідно враховувати всі рівні абстракції даних лексикографічних систем, що інтегруються: здійснювати аналіз концептуальних моделей, способів подання даних на внутрішньому рівні, архітектур реалізації зовнішнього рівня, щоб визначити, на якому рівні можна здійснити інтеграцію, чи, можливо, цей процес охоплюватиме два або й усі три вищезазначені рівні. Інколи цей аналіз демонструє неможливість інтеграції вихідних інформаційних систем, тому завдання змінюється на розроблення зовсім нового об'єкта. На наступному етапі завдання зводиться до завдання інтеграції інформаційних систем, тому перед розробником постає потреба вибору типу інтеграції відповідно до архітектур, які покладено в основу кожного з компонентів інтеграції.

Завдання інтеграції об'єктів нашого дослідження потребує подальшого опрацювання, проте на основі отриманих даних можна зробити висновки, що цей процес не ускладнений використанням різних підходів до проектування компонентів інтеграції, а отже, з технічного погляду інтеграція можлива на будь-якому із трьох рівнів структури лексикографічних систем. Тип інтеграції, а також рівнів, які вона охоплюватиме, буде вибрано після аналізу концептуальних моделей лексикографічних систем Словника української мови, Етимологічного словника української мови та Граматичного словника української мови. Основні аспекти та результати цього процесу будуть висвітлені в наступних роботах.

1. Steel T. Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems 75-02-08 / T. Steel // ACM SIGMOD Record / T. Steel. – New York: ACM, 1975. – ISSN: 0163-5808. – T. 7. – No. 2. – С. 1–140. 2. Рабулець О. Г. Інтегровані лексикографічні системи : автореф. дис... канд. техн. наук: 05.13.06 / О. Г. Рабулець; НАН України. Нац. б-ка України ім. В. І. Вернадського. – К., 2002. – 18 с. 3. Словники України. Інтегрована лексикограф. система (версія 4.1) [Електронний ресурс] : словозміна, транскрипція, фразеологія, синонімія, антонімія / [В. А. Широков та ін.] ; Нац. акад. наук України, Укр. мовно-інф. фонд. – Електрон. дан. – К. : Довіра, 2010. – 1 електрон. оптич. диск. – Систем. требования: Операційна система – MICROSOFT WINDOWS 7/ MICROSOFT WINDOWS VISTA / MICROSOFT WINDOWS SERVER 2008 / MICROSOFT WINDOWS SERVER 2003 / MICROSOFT WINDOWS XP SP3 ; програмне забезпечення MICROSOFT, NET 4.1; процесор INTEL PENTIUM/CELERON/ XEON, AMD K6/ATXLON / DURON або сумісний з ними процесор, тактова частота якого 1 ГГЦ і вище; оперативна пам'ять – не менше 512 МБ ; вільне місце на жорсткому диска – не менше 100 МБ; дисковод для CD/DVD-дисків – для встановлення програми. – Назва з етикетки диска. – ISBN 978-966-507-275-1. 4. Широков В. А. Комп'ютерна лексикографія / В. А. Широков. – К.: Наук. думка, 2011. – 351 с.