УДК 618.3.01

О.В. КАСИЛОВ

МОДЕЛИРОВАНИЕ СЛОВАРЯ-ТЕЗАУРУСА

У статті наведені методи формалізації структури паперового словника-тезауруса, при цьому словник розглядається як різновид інформаційної системи. Пропонується набір правил перетворення структури паперового словника в його електронну форму з використанням мови розмітки структурованих текстів XML.

In this article the formalization methods of paper dictionary - thesaurus structure is listed, in this connection the dictionary is considered as a version of information system. The procedures set that transform paper dictionary structure to its electronic form using sectoring of structured texts XML language is offered.

Постановка проблемы. В настоящее время создание корпусов текстов, разработка лингвистических баз данных, информационно-поисковых систем – это приоритетное направление в компьютерной лингвистике. Создание электронных словарей-тезаурусов является одной из задач. Формализация структуры бумажного словаря-тезауруса, и разработка правил преобразования словарной статьи словаря в электронную форму позволяет разработать лингвистический процессор, автоматизирующий процесс создания электронных версий словарей-тезаурусов, являющихся основой лингвистических баз данных.

Анализ литературных источников показал, что разработки в области технологий создания электронных словарей ведутся не высокими темпами. Как указано в «Каталоге лингвистических программ и ресурсов в Сети» доступны только 4 электронных версии тезаурусов для английского языка, а для украинского и русского языков такие разработки отсутствуют.

Разработки, проводимые в последнее время, в основном сосредоточены на пополнении существующего ядра лингвистической базы данных [1 – 3], проблем формализации в лингвистике [4], а не на разработку лингвистического процессора, позволяющего автоматизировать процесс создания электронного словаря. Ряд авторов обратили внимание на необходимость использования специализированных языков разметки таких как XML [5, 6], что подтверждает правильность выбора этого средства для работы со структурированными данными, которыми являются бумажные словари.

Цель статьи – анализ логической структуры бумажного словаря-тезауруса, формализация его структуры и разработка правил преобразования словарной статьи словаря-тезауруса в электронную форму.

Основной раздел. Словарь как абстрактная лексикографическая система обязательно имеет структуру, которая содержит две части: левая (реестровая) и правая (интерпретационная).

Именно наличие правой части отличает словарь от списка слов. Но словарь имеет и более глубокую структуру, которая отображается в строении левых и правых частей словаря в целом и его словарных статей, а также в структуре межстатейних и межсловарных отображений.

Таким образом, словарь представляет собой специальный вид текста, в котором в систематизированном и структурированном виде представлено описание лексики определенного языка (или совокупности языков). Однако словарь рассматривается и как специфический объект техники, а именно – информационную систему, где посредством полиграфического исполнения обозначаются лингвистические эффекты с помощью шрифтовых выделений, позиционного размещения, специальных пометок и т.п., которые играют роль идентификаторов соответствующих информационных переменных.

Сложность построения словаря состоит в том, что не все элементы его структуры явным образом обозначены указанным выше способом. В структуре реальных словарей, как правило, большое количество неявных структурных элементов, выявление которых часто является довольно сложной задачей.

Множество структурированных элементов словаря вместе со способами их комбинирования составляют своеобразный метаязык словаря, определения системных характеристик которого может быть основой для развития соответствующих формальных моделей. Процесс абстрагирования словарного метаязыка является специфической разновидностью лексикографического эффекта.

Информационно-лексикографическая модель любой лексикографической системы (или ее реализация) может быть представлена в следующем виде:

$$V(\ell) = \{\Lambda(\ell), P(\ell), H\}, \qquad (1)$$

где $V(\ell)$ – лексикографическая система как множество словарных статей; $\Lambda(\ell)$ – множество левых частей словарных статей словаря $V(\ell)$; $P(\ell)$ – множество правых частей этого же словаря; H – отображение множества $\Lambda(\ell)$ на $P(\ell)$:

$$H : \Lambda(\ell) \to P(\ell). \qquad (2)$$

В определении лексикографической системы отображение H выступает функцией, которая ставит в соответствие левой части словарной статьи ее правую часть и обеспечивает дихотомическую целостность построения соответствующей словарной статьи.

Лексикографическое размежевание левой и правой частей касается не столько формально позиционного их расположения в словарной статье, сколько отображения функционального противопоставления формы и

содержания в слове. В бумажных словарях в их непосредственном печатном исполнении часты случаи «перемешивания» отдельных элементов структуры $\Lambda(\ell)$ на $P(\ell)$.

В структуре лексикографической модели, как и в бумажных словарях, словарная статья начинается соответствующим заголовочным словом *x*. Поэтому, формула (1) записи словарных статей лексикографической системы детализируется таким образом:

$$V(\ell)=\underset{x\in S_0(\ell)}{\cup}V(x),\ \Lambda(\ell)=\underset{x\in S_0(\ell)}{\cup}\Lambda(x),\ P(\ell)=\underset{x\in S_0(\ell)}{\cup}P(x), \qquad (3)$$

где $V(x)$ – словарная статья лексикографической системы $V(\ell)$, начинающаяся заголовочным словом *x*;

$\Lambda(x)$ – левая часть словарной статьи; $P(x)$ – правая часть словарной статьи; $V(x)$ – однозначная функция от *x* и заголовочное слово *x* выступают как идентификатор $V(x)$.

Из определения отображения *H* следует, что $H(\Lambda(\ell)=P(\ell))$, причем функция *H* является однозначной. На множестве $V(\ell)$ определяется частичный порядок, индуцированный «лексикографическим» упорядочением множества $S_0(\ell)$.

Среди основных структурообразующих элементов лексикографической системы $V(\ell)$ выделяем систему ее автоморфизмов, то есть отображений

$$A:\Lambda(\ell)\to V(\ell) \qquad (4)$$

системы $V(\ell)$ в себя. Автоморфизм *A* может, в частности, констатировать наличие отсылочных типов словарных статей, например, таких: *x* см. *y*. Указанный автоморфизм определяет такое отображение словарных статей: $V(x)\to V(y)$. Его идентификатором является, как правило, некоторое отсылочное псевдослово (в приведенном примере – см. *y*), которое противопоставляет словарной статье $V(x)$ ее соответствие $V(y)$. Заметим, что строение автоморфизма А может быть более сложным, чем в этом примере.

Во-первых, длина ряда отсылок может быть больше единицы, то есть иметь рекурсивный характер:

$$V(x)\to\{V(x')\}\to\ldots\to\{V(x''')\}\to\ldots\ .$$

Кроме того, отображение $V(x)\to V(y)$ может репрезентовать целый пучок отсылок. Это реализуется, если словарная статья $V(x)$ имеет такое строение:

x, x', x'', ... см. y, y', y'', ...

В этом случае в одной словарной статье $V(x)$ определен пучок отображений:

$$V(x)\to V(y);\ V(x')\to V(y'),\ V(x'')\to V(y'')\ \ldots\ .$$

Для формирования лексикографической базы данных необходимо выполнить формализацию структуры словаря; с этой целью введем дополнительные обозначения.

Обозначим через $V(x)$ словарную статью с реестровой единицей X (дескриптор); $\Lambda(x)$ часть словарной статьи, которая содержит заголовочный ряд; через $S(x)$ – часть словарной статьи, которая обозначает парадигматические отношения дескриптора и состоит из отдельных значений, содержащих ряд объяснений и помет S_i, заголовочного слова (термина) X:

$$S(X) = \bigcup_{i=1}^{n} S_i . \quad (5)$$

Через F_{ij} обозначим часть словарной статьи, которая включает группу ее терминологических словосочетаний, которые входят в состав значения S_i:

$$S_i = \bigcup_{j=1}^{n} F_{ij} . \quad (6)$$

Таким образом:

$$V(X) = \Lambda(X) \cup S(X) . \quad (7)$$

Если принять соглашение:

$$A \subset B \Leftrightarrow A \leftarrow B ,$$

то каждой словарной статье можно поставить в соответствие графический представитель ее структуры.

Элементарной структурной единицей тезауруса является словарная статья дескриптора, которая строится по алфавитно-структурному принципу

$$d_i < M_{i1}, M_{i2}, M_{i3}, M_{i4} > ,$$

где d_i – заглавный дескриптор; M_{i1} – упорядоченное по алфавиту множество условных синонимов данного заглавного дескриптора, образующих вместе с ним класс условной эквивалентности; M_{i2} – упорядоченное по алфавиту множество дескрипторов, каждый из которых связан с заглавным дескриптором отношением «род – вид»; M_{i3} – упорядоченное по алфавиту множество дескрипторов, каждый из которых связан с заглавным дескриптором отношением «вид – род»; M_{i4} – упорядоченное по алфавиту множество дескрипторов, каждый из которых связан с заглавным дескриптором по крайней мере одним из следующих парадигматических отношений: целое – часть, часть – целое, причина – следствие, следствие – причина, функциональное сходство (ассоциативные связи).

Любое из перечисленных множеств может быть одноэлементным и даже пустым, т.е. может отсутствовать в словарной статье.

Множество M_{i1} в совокупности с дескрипторами d_i образуют класс условной эквивалентности, который также является дескриптором. Это

множество M_{i1} выполняет функцию номинального определения, которое уточняет смысл дескриптора d_i, выбранного для обозначения этого класса условной эквивалентности.

Рассмотрев структуру словарной статьи словаря-тезауруса [7] и ее запись на языке разметки структурированных текстов XML [8], запишем набор правил для преобразования входных данных (словарная статья) в выходные данные (словарная статья в XML записи).

$$\left.\begin{array}{l} \text{ПР1}\left(T_0^j\right) = R1, \\ \text{ПР2}\left(T_1^j\right) = R2, \\ \text{ПР2}\left(T_2^j\right) = R3, \\ \text{ПР4}\left(T_3^j\right) = R4, \\ \text{ПР5}\left(T_4^j\right) = R5 \end{array}\right\} j = 1,\overline{N}, \qquad (8)$$

где T_0 – дескрипторная группа; T_1 – родовой дескриптор; T_2 – видовой дескриптор; T_3 – дескриптор; T_4 – условный синоним; T_5 – ассоциативный дескриптор; j – словарные статьи словаря; ПР_n – программы(а), выполняющие преобразование; R_n – результат преобразования.

Рассмотрим абстрактный пример словарной статьи, которую запишем в символьной форме и укажем соответствующие правила преобразования (8) к входному потоку данных:

Обозначение	Символьная форма (Входные данные)	Преобразование	Выходные данные
dg_i	НГ ИГ	ПР1	<area id = “НГ” name = “ИГ”/>
d_i	Д НГ (П)	ПР2	<descript id = “Д” area = “НГ”/> </descript> <explan>(П) </explan>
M_{i1}	ИВ Див	ПР3	<syn id = “Див”/>
M_{i2}	РД Дрд	ПР4	<child id = “Двд”/>
M_{i3}	ВД Двд	ПР5	<child id = “Двд”/>
M_{i4}	АД Дад	ПР6	<assoc id = “Дад”/>

Выводы. Анализ логической структуры бумажного словаря-тезауруса позволил провести формализацию его структуры и разработать правила

преобразования словарной статьи словаря-тезауруса в электронную форму с использованием языка разметки структурированных текстов XML, предложена система дескрипторов на базе правил XML для представления словарной статьи в электронной форме.

Разработанная модель словаря-тезауруса и набор правил преобразования словарной статьи словаря-тезауруса в электронную форму являются основой формирования лексикографического процессора. При формировании лексикографической базы данных структурообразующие элементы играют роль элементов структуры базы данных и ее поисковых параметров. Формирование лексикографической базы данных после проведенной формализации структурообразующих элементов словаря редуцируется к полностью автоматической процедуре, которой подвергается текст.

Словарь может быть использован как фрагмент лингвистического обеспечения создаваемых автоматизированных систем, связанных с соответствующей предметной областью.

Список литературы: 1. *Герд А.С.* Базы данных и прикладная лингвистика. Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под. ред. А.С. Герда. – СПб.: Изд-во С.-Петерб. ун-та, 2002. – 168 с. **2.** *Азарова И.В. и др.* Разработка компьютерного тезауруса русского языка типа WordNet. Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под. ред. А.С. Герда. – СПб.: Изд-во С.-Петерб. ун-та, 2002. – 168 с. **3.** *Azarova L., and others*. RussNet: Building a Lexical Database for the Russian Language // Proceedings of Workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation in LREC2002, June 2002. Las Palmas de Gran Canaria, 2002. **4.** *Широков В.А., Рубанець О.Г.* Формалізація у галузі лінгвістики // Актуальні проблеми української лінгвістики: теорія і практика. – К., 2002. – Вип. 5. С. 3-27. **5.** *Андреев А.В.* Представление данных в Индоевропейском компьютерном тезаурусе (ThIE). Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под. ред. А.С. Герда. – СПб.: Изд-во С.-Петерб. ун-та, 2002. – 168 с. **6.** *Boguslavsky I.M., Grigorieva S.A., Grigoriev N.V., Kreidlin L.G., Frid N.E. Dependency* Treebank for Russian: Concepts, Tools, Types of Information // Proceedings of the 18th Conference on Computational Linguistics. Vol. 2. Saarbruecken, 2000. P. 987-991. **7.** *Касилов О.В.* Методы представления структурированных текстов естественного языка в XML описании // Вісник НТУ «ХПІ». Збірка наукових праць. Тематичний випуск: Нові рішення у сучасних технологіях. – Харків: НТУ «ХПІ». – 2002. – № 6. – Т. 2 – 156 с. **8.** *Касилов О.В., Самойлов А.Н., Шраер А.С.* Основы разметки текстов // Вісник НТУ «ХПІ». Збірка наукових праць. Тематичний випуск: Автоматика та приладобудування. – Харків: НТУ «ХПІ». – 2002. – № 9. – Т. 7. – 198 с.

Поступило в редакцию 20.04.04