Лексикографическая структура этимологического словаря и его представление в цифровой среде

Etymological dictionary: lexicographic structure and representation in digital environment

Остапова И. В. (iros@zeos.net)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

На основе формальной модели лексикографической системы Этимологического словаря украинского языка разработана технология построения инструментальной системы для поддержки функционирования словаря в цифровой среде. Основное внимание уделяется механизму языковой индексации словаря.

Этимологический словарь украинского языка (далее — ЭСУЯ), представляет собой фундаментальный лексикографический труд, который создаётся в рамках проекта формирования национальной словарной базы Украины [1]. Первый том был издан в 1982 году, выход шестого тома ожидается в 2009 году; седьмой том должен будет представлять индекс ко всему словарному массиву. Для этимологических словарей наиболее действенным поисковым инструментом является индекс по языковой принадлежности слов, с которыми устанавливается генетическая связь в каждой словарной статье.

На сегодняшний день (на материале 5-ти вышедших из печати томов) представлены уже 232 различных языка. Для каждого языка в словаре необходимо создать отдельный индекс с идентификацией всех точных текстовых локализаций каждого слова этого языка. Для всего массива словаря предполагаемая размерность индекса — около 120 тысяч единиц. Трудоемкость работы по построению такого индекса столь велика, что задача создания его в «ручном» режиме не представляется технологически оправданной. Поэтому была поставлена задача разработки специальной цифровой лексикографической среды, адаптированной к структурам ЭСУЯ и ориентированной на создание моногоязыкового индекса в автоматическом режиме.

Цифровая среда представляет собой качественно новый уровень сервиса для исследовательской работы с лингвистической информацией, представленной в словарной форме. И в первую очередь это относится к индексным системам. Под индексацией словаря мы понимаем набор формализованных правил и процедур, на основании которых можно получить информацию об определённых языковых фактах, зафиксированных в словаре. Реализуются эти правила в форме пользовательских интерфейсов. Однако следует учитывать тот факт, что эффективность автоматического построения индексных схем для цифрового словаря возможна только в достаточно формализованной среде.

При выполнении работы по созданию цифровой версии ЭСУЯ использовались методы, которые уже был успешно опробованы для решения подобных задач, в частности, для создания компьютерной лексикографической базы данных нового толкового Словаря украинского языка [3].

Мы рассматриваем словарь как информационную систему особого типа — лексикографическую. Согласно теории лексикографических систем это абстрактный языково-информационный объект, ориентированный на реализацию комплексного информационного описания лексико-грамматических структур определённого языка или совокупности языков [3].

Архитектура системы отвечает стандартной трёхуровневой архитектуре информационных систем ANSI/SPARK, согласно которой в информационной системе выделяются концептуальный, внутренний и внешний уровни данных [2].

В качестве концептуальной модели используется лексикографическая модель данных [3]. Ниже мы приводим её в несколько упрощённом виде:

$$\{I_0(D),\ V(I_0(D)),\ \beta,\ \delta[\beta],\ Red[V(I^Q(D))]\},$$

где D — объект моделирования — Этимологический словарь украинского языка; $I_0(D) = \{x_1\}$ множество

реестровых единиц словаря, в теории лексикографических систем его принято называть множеством *элементарных информационных единиц*; $V(I_0(D))$ — множество описаний (интерпретаций) элементарных информационных единиц, то есть текстов словарных статей: $V(I_0(D)) = \{V(x_i)\}$ — словарная статья с заголовковым словом (реестровой единицей) x_i; β — множество структурных элементов, которые были абстрагированы в результате анализа текста словаря; δ[β] — структура, которая порождается на β оператором δ; ограничения δ[β] на $V(x)$ порождает микроструктуру словарной статьи δ(x); $Red[V(I_0(D)]$ — механизм *рекурсивной редукции лексикографической системы*. Он даёт возможность последовательно выявлять всё более тонкие детали структуры лексикографической системы, в частности — осуществлять распределение структурных элементов словарной статьи на реестровую и интерпретационную части.

Концептуальная модель словаря строится на основе анализа полиграфической версии ЭСУЯ, то есть анализируется типографское оформление, организация и структура печатных текстов словарных статей, которые интерпретируются как идентификаторы соответствующих элементов лексикографических структур β и $\delta(x)$.

В качестве базового структурного элемента лексикографической системы ЭСУЯ мы определяем *этимологический класс*, который представляет собой блок линейного текста словарной статьи, в котором описаны определённые генетические связи реестрового украинского слова. Вычленение этимологических классов выполняется по формальным признакам: структурная единица идентифицируется как этимологический класс, если в тексте словарной статьи можно выявить уникальные знаковые последовательности, используемые в качестве разделителей. Для ЭСУЯ нами выделены следующие типы этимологических классов: *класс реестрового слова* (обозначим *HEAD*), *класс дериватов* (*DER*), *класс славянских соответствий* (*SLAV*), *языковой класс* (*LANG*), *библиографический класс* (*BIBL*), *класс ссылок* (*LINK*). Отметим, что тип описания «языковой класс» используется только как наименование структурного элемента концептуальной модели словаря, а не как лингвистический термин. Каждый из этих классов имеет уникальную структуру текста, что дало нам возможность построить процедуру идентификации типа каждого этимологического класса в словарной статье по формальным признакам.

Дадим краткую характеристику каждого класса. *Класс реестрового слова* содержит собственно реестровое (заголовковое) слово и определённые его параметры. Заголовковым может быть слово как литературное, так и диалектное, а также имя собственное. Этот класс уникальный и обязательно входит в состав словарной статьи.

К *языковому классу* относим следующие описания: а) реконструируемые формы реестрового слова или их основы на разных этапах развития праславянского языка, представленные в антихронологическом порядке; б) этимологически связанные с реестровым словом слова других индоевропейских языков, начиная с ближайших к праславянскому фонетических и словообразовательных форм; в) этимологически связанные с реестровым словом слова семито-хамитских или урало-алтайских языков; г) этимологическая связь слова не установлена, например «этимология неясная». Анализ ЭСУЯ показал, что таких классов максимум два в словарной статье, но мы не ограничиваем их количество в нашей модели. В состав словарной статьи должен входить хотя бы один класс данного типа.

Класс реестрового слова и *языковый класс* составляют минимальную структуру словарной статьи. Этимологические классы других типов являются факультативными.

Класс дериватов содержит родственные с реестровым словом слова украинского языка, то есть ближайшие этимологически значения. В тексте словарной статьи может быть не более одного этимологического класса этого типа.

Класс славянских соответствий содержит соответствия реестрового слова из всех славянских языков, в которых они зафиксированы. В словарной статье может быть не более одной структурной единицы этого типа.

Библиографический класс — блок текста, содержащий информацию о научных трудах, в которых рассматривается этимология соответствующего украинского слова или связанных с ним слов других языков. Такой класс может быть только один.

К *классу ссылок* относим те текстовые блоки, где описываются связи с другими статьями словаря.

Проиллюстрируем сказанное на примере двух небольших, но достаточно репрезентативных с точки зрения структуры словарных статей. Тексты приводим в форме, максимально приближенной к печатной версии.

*Пример 1 (словарная статья с заголовковым словом **абетка**):*

абетка, [*абетло*] Пі, *абетний* (заст.) «елементарний»**;—** власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука*, *альфабет* і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)**.—** Sadn. — Aitz. VWb. I 42**.—** Пор. **азбука, алфавіт.**

*Пример 2 (словарная статья с заголовковым словом **абзац**):*

абзац;— р. бр. *абзац*, болг. *абзац*, схв. *абзац*;**—** запозичення з німецької мови**;** нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan,

англ. set і спорідненого з псл. saditi, укр. *садити*. — CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705. — Див. ще **абажур, садити.** — Пор. **обцас.**

Пример 3 (этимологические классы для словарной статьи ***абетка;*** *тексты классов подаются в угловых скобках):*

HEAD ≡ <**абетка**>
DER ≡ <[*абетло*] Пі, *абетний* (заст.) «елементарний»>
LANG ≡ <власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука*, *альфабет* і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)>
BIBL ≡ <Sadn. — Aitz. VWb. I 42>
LINK ≡ <Пор. **азбука, алфавіт**>

Пример 4 (этимологические классы для словарной статьи ***абзац****):*

HEAD ≡ <**абзац>**
SLAVIA ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>
LANG ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *садити*>
BIBL ≡ <CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705>
$LINK_1$ ≡ <Див. ще **абажур, садити>**
$LINK_2$ ≡ <Пор. **обцас>**

Приведём пример словарной статьи, минимальной как структурно, так и содержательно:

Пример 5 (этимологические классы для словарной статьи ***андріяк****):*

[андріяк] «опій»; — походження неясне.
HEAD ≡ <**[андріяк]** «опій»**>**
LANG ≡ <походження неясне>

В тексте каждого этимологического класса устанавливаются связи реестрового слова с определёнными словами других языков. Все эти слова, включая реестровые, мы будем называть *этимонами.* При анализе текстов этимологических классов было выявлено восемь параметров, посредством которых описываются этимоны: *маркер языковой принадлежности* (обозначим P_L), *ремарка к маркеру языковой прнадлежности* (P_{RL}), *знаковое представление этимона* (P_A), *принадлежность к диалектной лексике* (P_D), *маркер омонимии* (P_o), *толкование* (P_S), *ремарка* (P_R), *библиография* (P_B). Мы перечислили параметры в том порядке, в котором они, как правило, следуют в тексте соответствующего этимологического класса. Два параметра являются обязательными: это P_L (*маркер языковой принадлежности*) и P_A (*знаковое представление этимона*). Эти два параметра обеспечивают уникальность каждого этимона словарной статьи: этимоны с одинаковой знаковой формой могут иметь разную языковую принадлежность, или этимоны с одинаковой языковой принадлежностью могут иметь разные знаковые формы. Остальные параметры — факультативные. Для каждого параметра определена формальная процедура, которая позволяет вычленить соответствующий параметр из текста для каждого этимологического класса.

Набор параметров $\{P_L, P_{RL}, P_A, P_D, P_O, P_S, P_R, P_B\}$ мы будем называть *этимон-структурой* и будем обозначать символом $ETYM(e_i)$, где e_i — соответствующий этимон; индекс *i* — порядковый номер данного этимона в тексте. Порядок следования параметров в этимон-структуре полагаем не существенным.

Не все параметры актуальны для каждого этимологического класса. Текст, который мы идентифицируем как этимологический класс, использует свое подмножество параметров; не каждый этимон обязан описываться полным набором параметров. Однако для достижения структурной однородности для каждого класса строится один тип этимон-структуры; если определённый параметр не задействован или не может быть выделен по формальным признакам, то его значению соответствует пустая строка текста. Этимон-структура строится только в том случае, если удалось вычленить P_A. Формально мы полагаем, что каждому этимологическому классу соответствует этимон-структура. Если языковой класс не имеет ни одного этимона (или не удалось его выявить формальной процедурой), то мы считаем его вырожденным этимологическим классом и ему соответствует пустая этимон-структура. Примером такого класса служит языковой класс для словарной статьи из примера 5.

Проиллюстрируем этимон-структуры на примерах текстов этимологических классов:

Пример 6 (этимон-структуры для класса реестрового слова):

HEAD (**абзац**) ≡ <абзац**>**
ETYM (e_1) ≡ $\{P_L$ = <укр.>, P_A = <абзац>$\}$

Пример 7 (этимон-структуры для класса дериватов):

DER (**абетка**) ≡ <[*абетло*] Пі, *абетний* (заст.) «елементарний»>
ETYM (e_1) ≡ $\{P_L$ = <укр.>, P_A = <абетло>, $P_D = 1, P_B$ = <Пі>$\}$
ETYM (e_2) ≡ $\{P_L$ = <укр.>, P_A = <абетний>, P_R = <(заст.)>, P_S = <«елементарний»>$\}$

Параметр омонимии P_O для этимона e_1 принимает значение 1, так как квадратные скобки указывают на принадлежность слова к диалектной лексике. По умолчанию для всех этимонов значение этого параметра 0.

Пример 8 (этимон-структуры для класса славянских соответствий):

SLAV (**абзац**) ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>

ETYM (e_1) ≡ {P_L= <р.>, P_A = <абзац>}

ETYM e_2) ≡ {P_L= <бр.>, P_A = <абзац>}

ETYM (e_3) ≡ {P_L= <болг.>, P_A = <абзац>}

ETYM (e_4) ≡ {P_L= <схв.>, P_A = <абзац>}

Пример 9 (этимон-структуры для языкового класса):

LANG (**абзац**) ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *садити*>

ETYM (e_1) ≡ {P_L= <нім.>, P_A = <Absatz>, P_S = <«перерва, пауза, уступ, абзац»>}

ETYM (e_2)≡ {P_L= <нім.>, P_A = <absetzen>, P_S = <«відсувати, відставляти»>}

ETYM (e_3) ≡ {P_L= <нім.>, P_A = <ab->, P_S = <«від-, з-»>}

ETYM (e_4) ≡ {P_L= <гот.>, P_A = <af>, P_S = <«від»>}

ETYM (e_5)≡ {P_L= <лат.>, P_A = <ab>, P_S = <«тс.»>

ETYM (e_6)≡ {P_L= <нім.>, P_A = <setzen>, P_S = <«тс.»>}

ETYM (e_7)≡ {P_L= <двн.>, P_A = <sezzen>}

ETYM (e_8) ≡ {P_L= <дангл.>, P_A = <settan>}

ETYM (e_9) ≡ {P_L= <англ.>, P_A = <set>}

ETYM (e_{10}) ≡ {P_L= <псл.>, P_A = <saditi>}

ETYM (e_{11}) ≡ {P_L= <укр.>, P_A = < садити >}

Основная проблема создания компьютерных словарей, исходя из их печатных версий, — это формирование соответствующей базы данных в автоматическом режиме непосредственно из текста словаря (парсинг). Опыт убеждает, что формирование лексикографических баз данных «вручную» из больших и сложных словарных текстов практически невозможно. Основная задача парсинга — автоматическое выделение определенных нами структурных элементов непосредственно из текста словаря, поскольку именно они выполняют роль элементов лексикографической базы данных.

Перед конверсией тексты всех томов были переведены в формат HTML и унифицированы как относительно структуры файлов, так и относительно знаковой системы. Словарь был подготовлен к печати различными издательскими технологиями. Первые три тома — в технологии монотайп, докомпьютерной. Поэтому печатные тексты сначала были отсканированы, распознаны программой FINEREADER, а затем вычитаны. Тексты 4-го и 5-го томов уже готовились в издательской системе, т. е. в цифровом формате. Знаковая система всех текстов словаря была унифицирована согласна кодировке UNICODE 3.0. Это позволяет выполнить инвентаризацию символов алфавита для представления этимонов каждого языка.

В результате этих операций были получены специальным образом препарированные тексты томов Этимологического словаря, полностью готовые для автоматической конверсии в лексикографическую базу данных.

Для поддержки цифровой версии словаря построен инструментальный комлекс, который обеспечивает такие основные функции [4]:

1) автоматическую конверсию текстов этимологического словаря в компьютерную базу данных;
2) традиционный вход в систему по реестровому слову и отображение текста словарной статьи;
3) редактирование любого структурного элемента словарной статьи;
4) построение этимон-структуры для словарной статьи в ручном режиме;
5) автоматическое построение этимон-структуры для словарной статьи;
6) создание словарной статьи с определенной структурой.

На рис. 1 показано одно из окон редактирования словарной статьи. На левой панели словарная статья представлена в виде дерева структурных элементов. Для каждого этимологического класса выводится упорядоченный список этимонов, тем самым графическими средствами визуализируется глубина этимологического исследования. С помощью кнопок на средней панели структурные элементы можно добавлять, удалять и менять порядок их следования. Функции кнопок варьируются в зависимости от выбранного структурного элемента. Так, например, кнопка «Додати» (добавить) при выборе этимона позволяет добавить только этимон. Для каждого структурного элемента разработана своё окно редактирования, которое отражает специфику этого элемента. Для каждого этимона выводится и текст соответствующего этимологического класса, однако с запретом его редактирования. Это даёт возможность верифицировать параметризацию этимона, выполненную автоматически.

Для автоматического построения языковых индексов разработано специальный инструментарий, который позволяет:

1) в интерактивном режиме формировать любое количество языковых регистров на множестве всех языков словаря;
2) задавать спектры индексации, учитывая структуру словарной статьи.

На рис. 2 показано диалоговое окно пользователя для формирования языкового регистра.

Левая панель предназначена для выбора уже сформированных регистров в качестве неизменяемых шаблонов. Правая панель используется для редактирования существующих и формирования новых регистров.

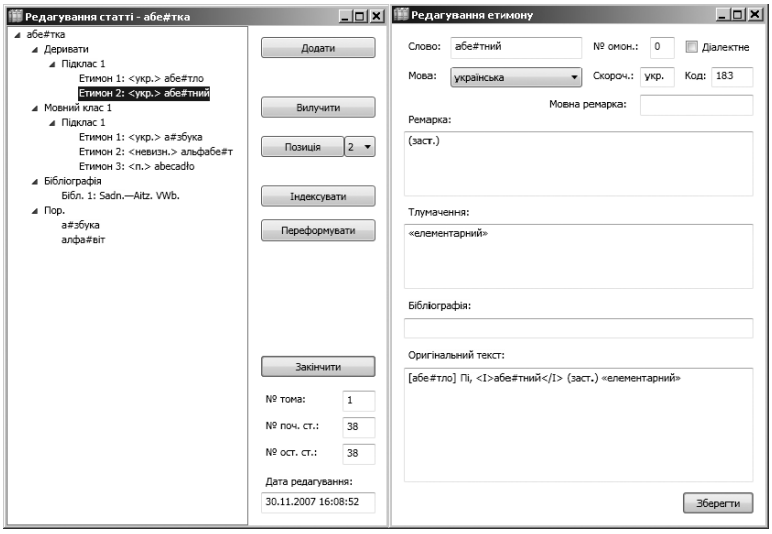

Рис. 1. Окно редактирования для словарной статьи с заголовковым словом ***абетка***

На рис. 3 показано окно главного пользовательского интерфейса словаря с индексом, построенным по сформированному регистру.

На левой панели из предложенного набора языков был выбран польский (возможен выбор как всех языков в регистре, так и определённого подмножества языков). На правой панели в реестровое окно выведен список всех этимонов, которые идентифицированы как слова польского языка.

В окно реестра могут быть также выведены заголовковые слова тех словарных статей, в которых зафиксированы этимологические связи с польским языком. Сформированный индекс по команде пользователя выводится в текстовый файл с указанием локализации каждого этимона.

Инструментальная система позволяет задать локализацию индексируемых элементов с точностью до структурного элемента — типа этимологического класса — словарной статьи (с помощью

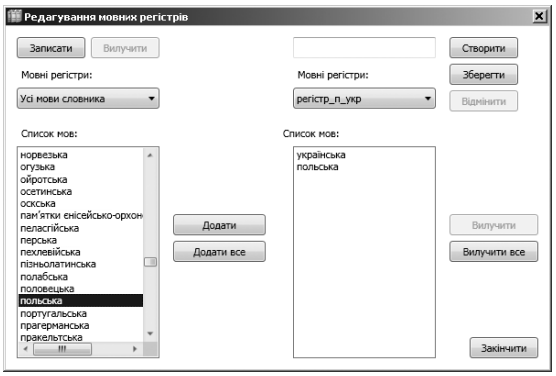

Рис. 2. Окно формирования языкового регистра

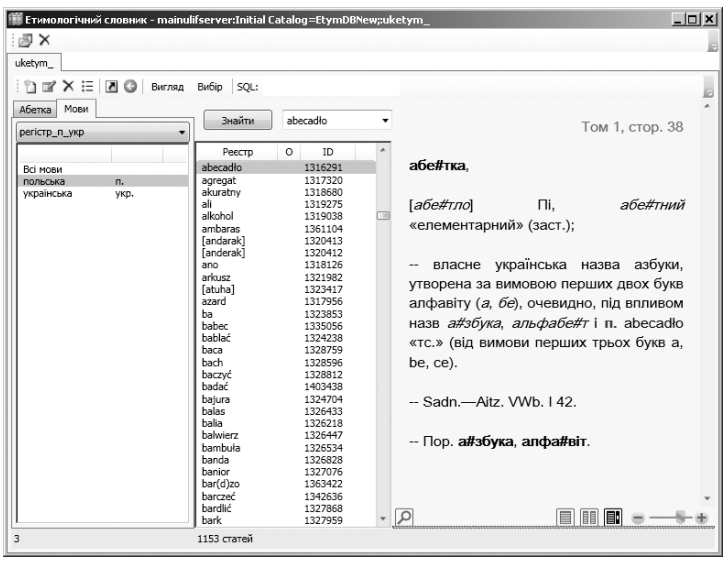

Рис. 3. Языковой индекс по заданному регистру

меню «Вибір» (Выбор) на верхней панели). В нашем случае был задан только *языковой класс*.

При активизации любого элемента реестра визуализируется текст словарной статьи.

Текст словарной статьи для вывода формируется из соответствующих полей базы данных. Полиграфическое оформление статьи практически сохранено полностью.

Описанный метод представления этимологического словаря в цифровой среде дал возможность построить для совокупности словарных статей словаря соответствующую совокупность этимон-структур как формальных репрезентантов описаний генетических связей реестровых единиц. Все схемы индексации строятся только на основе этимон-структур. Такой подход обеспечивает как возможность построения структур, имплицированных в текст словарных статей, так и отображения на цифровую среду аутентичного текста словаря, что делает цифровой словарь открытым для дальнейших интерпретаций.

Разработанные технологии и интерфейсы предлагаются как базовые для цифровых репрезентаций этимологических работ.

Литература

1. *Етимологічний* словник української мови: В 7 т. Київ: Наукова думка, 1982–2006. Т. 1–5.
2. *ANSI/X3/SPARK DBMS study group* interim report. FDT-Bull. ACM SIGMOD. 1975. V. 7. № 2.
3. *Широков В. А.* Елементи лексикографії. Київ: Довіра, 2005.
4. *Остапова И. В., Якименко К. Н.* Инструментальная лексикографическая система Этимологического словаря украинского языка // Прикладна лінгвістика та лінгвістичні технології. Київ: Довіра, 2008. С. 276–291.