И. В. Остапова

ЭТИМОЛОГИЧЕСКИЙ СЛОВАРЬ В ЦИФРОВОЙ СРЕДЕ

Как и все фундаментальные лексикографические работы, этимологические словари постепенно переводятся в цифровой формат. В интернете наиболее доступен «Этимологический словарь русского языка» М. Фасмера, первая его цифровая версия была издана на CD-ROM еще в 2004 году. Однако все цифровые репрезентации подобных лексикографических систем являются не более чем разновидностями машиночитаемых текстов, отсутствуют возможности масштабирования и динамического индексирования.

В данной работе рассматриваются метод и технология представления в цифровой среде Этимологического словаря украинского языка (ЭСУЯ, в печатной версии – 6 томов), которые предлагаются как универсальные для любой этимологической лексикографической системы [1]. Разработана формальная модель словарной статьи, в которой структурные элементы разделяются на два класса: лингвистические (описательные) и структурообразующие; последние вычленяются по формальной процедуре из лингвистических. На основе этой модели построена схема компьютерной базы данных (в рассматриваемой реализации – реляционной) и технология парсинга текста словаря, что позволило в автоматическом режиме конвертировать текст. Для поддержки цифровой версии ЭСУЯ разработан инструментальный комплекс, в который входит подсистема индексирования словаря по любому сформированному пользователем языковому регистру.

При выполнении работы по созданию цифровой версии ЭСУЯ использовалась методы, которые уже был успешно опробованы для решения подобных задач, в частности, для создания компьютерной лексикографической базы данных нового Словаря украинского языка [2; 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 определенного языка или совокупности языков [2].

Архитектура системы отвечает стандартной трехуровневой архитектуре информационных систем ANSI/SPARK, согласно которой в информационной системе выделяются концептуальный, внутренний и внешний уровни данных [4].

В качестве концептуальной модели мы используем лексикографическую модель данных [3]. Ниже мы приводим ее в несколько упрощенном виде:

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

где D – объект моделирования – Этимологический словарь украинского языка; $I_0(D) = \{x_i\}$ 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 дает возможность последовательно выявлять все 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 словарной статьи определен *этимологический класс* (обозначается *ECL*), представляющий собой блок линейного текста, в котором описываются определённые генетические связи заголовочного слова. Для ЭСУЯ этимологические классы подразделяются следующим образом: *класс реестрового слова* (*HEAD*), *класс дериватов* (*DERIVAT*), *класс славянских соответствий* (*SLAVIA*), *языковый класс* (*LANG*), *библиографический класс* (*BIBL*) и *классы ссылок* (*REF* и *COMP* в зависимости от типа ссылки).

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 а, be, ce).— Sadn. — Aitz. VWb. I 42.— Пор. **азбука, алфавіт.**

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 а, be, ce)>

BIBL ≡ <Sadn. — Aitz. VWb. I 42>

LINK ≡ <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. sęzzen, дангл. sęttan, 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. **общас.**

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. sęzzen, дангл. sęttan, 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. **общас**>

В тексте каждого этимологического класса устанавливаются связи реестрового слова с определенными 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. 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. Остальные параметры – факультативны. Для каждого параметра определена формальная процедура, которая позволяет вычленить соответствующий параметр из текста для каждого этимологического класса.

Набор параметров мы будем называть *этимон-структурой* и будем обозначать символом $ETYM(e_i)$, где e_i – соответствующий этимон; индекс – порядковый номер данного этимона в тексте.

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. sęzzen, дангл. sęttan, англ. set і спорідненого з псл. saditi, укр. *садити*>

$ETYM\,(e_1) \equiv \{<P_L = <\text{нім.}>, P_A = <\text{Absatz}>\}$
$ETYM\,(e_2) \equiv \{<P_L = <\text{нім.}>, P_A = <\text{absetzen}>\}$
$ETYM\,(e_3) \equiv \{<P_L = <\text{нім.}>, P_A = <\text{ab-}>\}$
$ETYM\,(e_4) \equiv \{<P_L = <\text{гот.}>, P_A = <\text{af}>\}$
$ETYM\,(e_5) \equiv \{<P_L = <\text{лат.}>, P_A = <\text{ab}>\}$
$ETYM\,(e_6) \equiv \{<P_L = <\text{нім.}>, P_A = <\text{setzen}>\}$
$ETYM\,(e_7) \equiv \{<P_L = <\text{двн.}>, P_A = <\text{sęzzen}>\}$
$ETYM\,(e_8) \equiv \{<P_L = <\text{дангл.}>, P_A = <\text{sęttan}>\}$
$ETYM\,(e_9) \equiv \{<P_L = <\text{англ.}>, P_A = <\text{set}>\}$
$ETYM\,(e_{10}) \equiv \{<P_L = <\text{псл.}>, P_A = <\text{saditi}>\}$
$ETYM\,(e_{11}) \equiv \{<P_L = <\text{укр.}>, P_A = <\text{saditi}>\}$

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. Это позволило выполнить инвентаризацию символов алфавита для представления этимонов каждого языка.

Для поддержки цифровой версии словаря построен инструментальный комлекс, который обеспечивает такие основные функции [3]:

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;

7) индексирование словаря по заданным параметрам.

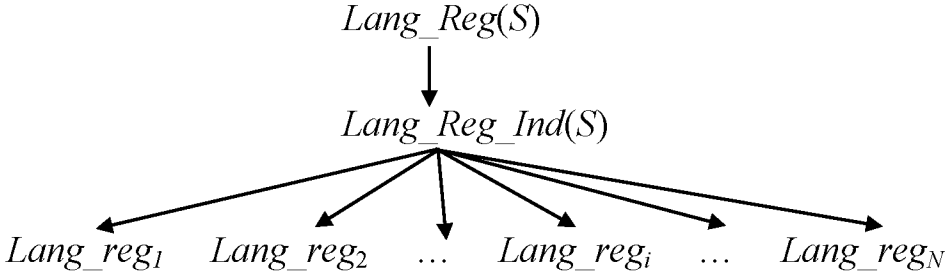
Одним из основных инструментов для эффективной работы со словарем является индекс. Индекс в его латентной форме присутствует в любом словаре: это заголовочные слова словарных статей, графически выделенные из массива текста и организованные в алфавитном порядке. При возрастании сложности структуры словарной статьи создание эффективной системы индексирования становится методологической и технологической проблемой, причем равно актуальной как для печатных словарей, так и для цифровых.

Исторически сложилось, что индексы разрабатывались, как правило, для двух типов словарей: фразеологических и этимологических. В первом случае необходимость индекса обусловлена структурной сложностью реестровой единицы; с помощью индекса организуется доступ к словарной статье по любой компоненте фразеологического словосочетания. Для этимологических словарей индекс является инструментом для установления генетической (эимологической) связи по параметру языковой принадлежности слова.

В идеале индекс должен быть построен по любому лексикографическому параметру словаря. Для реализации этого требования элементы индекса должны быть структурообразующими параметрами лексикографической системы, которая индексируется. Очевидно, развитые технологии индексирования могут быть построены только для цифровых лексикографических систем.

В предлагаемой модели структурообразующими являются обязательные параметры этимон-структуры: вход в словарь возможен по любому языку и по любому слову в алфавите этого языка.

Обобщенную схему многоязыкового индексирования этимологического словаря можно представить в таком виде:



Lang_Reg(*S*) – регистр всех языков (назовём его системным), наявных в этимологических описаниях словаря (к языкам также относим наречия и диалекты).

Lang_Reg_Ind(*S*) – языковой регистр, выбраный для индексирования словаря. Она может совпадать с *Lang_Reg_Ind*(*S*), или может быть его подмножеством: $Lang\_Reg\_Ind(S) \subseteq Lang\_Reg(S)$.

$Lang\_reg_i$ – некоторое подмножество индексного регистра. Как правило, в состав субрегистров входят родственные языки. Каждый субрегистр, в свою очередь, может подразделяться на более специализированные регистры ($Lang\_reg_{i_j}$).

При разработке инструментального комплекса языкового индексирования мы стремились минимально ограничивать пользователя системы. Разработанный инструментарий позволяет формировать языковой регистр с любым составом языков, зафиксированных в словаре, то есть объединять в одном регистре произвольно выбранные языки, не учитывая их родственность. Можна создать любое количество таких регистров. Регистры должны различаться только именем (названием), которое присваивается пользователем. Можно задать перечень структурных элементов (этимологических классов), которые будут индексироваться.

Регистр «СЛОВ'ЯНСЬКІ МОВИ» («славянские языки») на рис. 1 сформирован следующим образом: на основе системного регистра последовательно были сформированы «східно-слов'янські мови» («восточно-славянские языки»), «західно-слов'янські мови» (западно-славянские языки), «південно-слов'янські мови» («южно-славянские языки»), а затем языковые списки были объединены.

Возможность использование разработанного инструментария для других этимологических лексикографических систем в первую очередь обусловлено возможностью использования формальной модели. Модель строится поэтапно: анализируется общая структура словаря, его знаковая система, метаязык, выделяются тексты словарных статей из общего корпуса словаря, анализируется способ распределения статьи на этимологические классы и возможность распределения класов по типам, последний шаг – способ представления этимонов и их параметров в тексте и построение этимон-структур. Наши исследования показывают, что для любой этимологической систем вполне релевантна модель, разработанная для ЭСУЯ.

На базе инструментального комплекса построена виртуальная лексикографическая лаборатория – ВЛЛ «ЕСУМ» [2], которая обеспечивает профессиональное взаимодействие лексикографов в Интернет-среде и предоставляет возможность использования всего функционала системы в режиме on-line. Эволюция системы определяется четырьмя взаимосвязанными факторами: выделением более тонких структурных элементов из базовых структур, введением новых параметров словарной статьи, расширением функционала системы и разработкой новых интерфейсных схем.

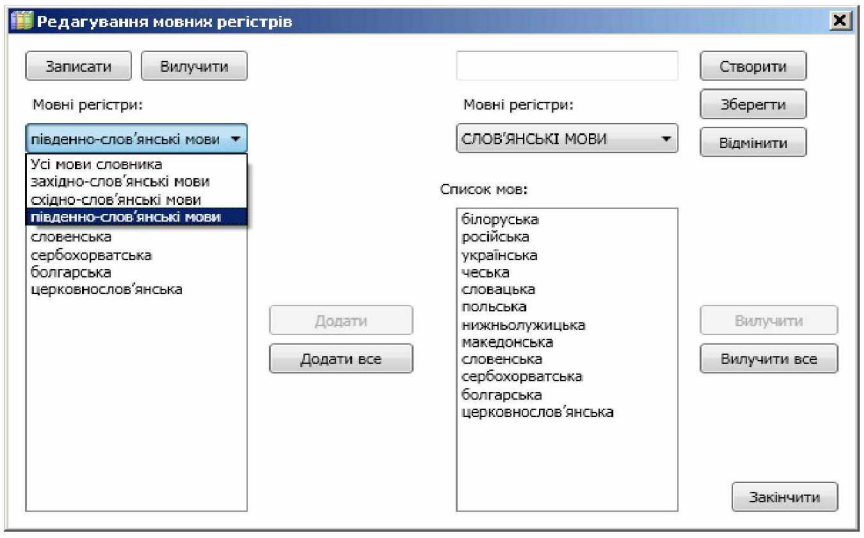


Рис. Формирование регистра «СЛОВ'ЯНСЬКІ МОВИ»

ЛИТЕРАТУРА

1. Етимологічний словник української мови. – Т. 1–6. – Київ : Наукова думка, 1982. – 2012.

2. *Остапова, И. В.* Виртуальная лексикографическая лаборатория для толковых словарей / И. В. Остапова, В. А. Широков // Компьютерная лингвистика и интеллектуальные технологии : матер. ежегод. Междунар. конф. «Диалог»; Бекасово, 26–30 мая 2010 г. – М. : РГГУ, 2010. – Вып. 9 (16). – С. 363–367.

3. СУМ: Словник української мови [Электронный ресурс]. – Режим доступа : http://lcorp.ulif.org.ua/ExplS. – Дата доступа : 12.03.2016.

4. *Остапова, И. В.* Лексикографическая структура этимологического словаря и его представление в цифровой среде / И. В. Остапова // Компьютерная лингвистика и интеллектуальные технологии : матер. ежегод. Междунар. конф. «Диалог 2009», Бекасово, 27–31 мая 2009 г. – М. : РГГУ, 2009. – Вып. 8 (15). – С. 359–365.