Виртуальная лексикографическая лаборатория для толковых словарей

Virtual lexicographical laboratory for explanatory dictionaries

Остапова И. В. (iros@zeos.net), **Широков В. А.** (vshirokov48@mail.ru)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

Обсуждаются принципы построения словарных систем в цифровой среде. Рассмотрена система толкового Словаря украинского языка в качестве примера интегрированной лексикографической системы, адаптирующей целый ряд лексикографических эффектов. Описана компьютерная сетевая технологическая среда, поддерживающая структуры толковых словарей, в форме виртуальной лексикографической лаборатории (ВЛЛ). ВЛЛ обеспечивает скоординированную работу территориально распределенного коллектива лингвистов, выполняющего масштабный лексикографический проект.

Очевидными преимуществами лексикографических трудов в цифровой форме являются практически неограниченный потенциал интеграции различных лингвистических фактов в едином объекте, способность к отражению языковой динамики, эффективность навигации по структурным элементам системы и возможность проведения вычислительных экспериментов. Это особенно важно для словарей большого объема, представленных в печатной форме многотомными изданиями. Немалое значение для ускорения перехода на компьютерные технологии лексикографирования имеет и предоставляемая цифровой средой возможность многократного использования однажды сформированных лексикографических структур и массивов многими профессионалами: лингвистами, лингвотехнологами и издателями. Особый смысл данная возможность приобретает в связи с развитием компьютерных коммуникаций.

В данной работе рассматриваются некоторые из отмеченных аспектов компьютерной лексикографии на примере проекта «Словарь украинского языка», представляющего одно из центральных заданий программы создания Национальной словарной базы Украины. В печатной версии новый словарь планируется в объеме 20 томов (в традиционном издательском формате многотомных толковых словарей). В лексикографическом плане Словарь представляет собой новую, существенно модифицированную версию созданного в период 1970–1980 гг. 11-томного толкового словаря украинского языка, в котором зафиксировано 134 тысячи слов [2]. Уже в процессе издания 11-томника возникла проблема его пополнения и модернизации. Новый Словарь должен максимально воспроизвести лексико-семантический состав украинского языка таким, каким он отражён в письменных источниках с конца 18 до начала 21 столетия, включая и источники Интернета. Основной корпус нового Словаря создан в течение 2002–2007 гг. В настоящее время завершена работа над первым томом, который уже передан в издательство. Статистические данные, сравнивающие первый том 20-томника (диапазон А–БЯЗЬ) с соответствующим диапазоном 11-томника, свидетельствуют о том, что фактически мы имеем новый словарный продукт. Это хорошо видно из таблицы:

Наименование показателя (диапазон А–БЯЗЬ)	СУЯ–11	СУЯ–20	Увеличение объема СУЯ-20 по сравнению с СУЯ-11 (%)
Словарные статьи	6303	11 527	82,88
Количество знаков в словарных статьях	1 786 334	3 922 154	119,56
Толкования	9577	14 334	49,67
Иллюстрации	11 604	26 388	127,40
Словосочетания, в том числе:	643	2249	249,77
– устойчивые словосочетания	439	520	18,45
– терминологические словосочетания	51	477	835,29
– эквиваленты слова	0	6	–
– фразеологизмы	153	1246	714,38

Общий объем текста диапазона в новом словаре увеличился более чем вдвое. Особенно значительно выросла подсистема словосочетаний — более чем в четыре раза, став по величине сравнимой с основной лексической частью. Данный факт стал побудительной причиной для более ясного изложения интегрированного представления о лексической и фразеологической семантике в лексикографической системе толкового словаря на всех уровнях ее архитектуры [3].

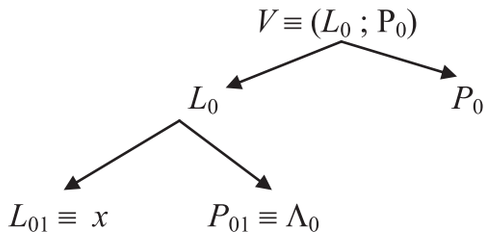
При создании нового Словаря был полностью использован не только опыт его предшественников, но и сам текст 11-томника в его полном объеме. С этой целью был произведен парсинг 11-томника и в автоматическом режиме сформирована его достаточно глубоко структурированная лексикографическая база данных, послужившая основой для создания виртуальной лексикографической лаборатории, описанию которой посвящена данная работа. К сожалению, авторам неизвестны примеры парсинга лексикографических продуктов такого масштаба и сложности, чем объясняется бедность ссылок на аналогичные работы [4].

В структуре Словаря украинского языка выделяем множество реестровых (заголовочных) слов: $W = \{x\}$, служащих идентификаторами соответствующих словарных статей $V(x)$. В структуре каждой словарной статьи $V(x)$ выделяется «левая часть» — $L(x)$, ответственная за описание грамматической семантики реестрового слова x, и «правая часть» — $P(x)$, в которой дается лексикографическое представление лексической семантики x. Кроме того, в системе определен оператор H: $L(x) \to P(x)$, обеспечивающий целостность словарной статьи и связь между лексической и грамматической семантикой (т. е. между грамматической формой и лексическим содержанием), которое несет лексема x, а также целый ряд других элементов (частью заданных неявно), отражающих те или иные аспекты лексикографического описания лексической системы.

В случае толкового словаря различаем два вида языковых единиц: единицы лексического уровня и словосочетания, которым в языке присвоен идиоматический статус. Поэтому естественно представить структуру словарной статьи $V(x)$ в виде объединения описаний структурных единиц обоих видов:

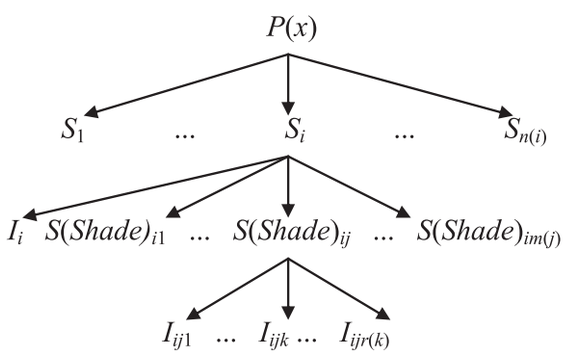
$$V(x) \equiv V^{Lex}(x) \cup [\bigcup_{i}^{n(x)} \bigcup_{j}^{m(i)} V^{j}{}_{i}{}^{Fras}(x)],$$

где $V^{Lex}(x)$ — описание семантики (грамматической и лексической) лексемы x; $V^{j}{}_{i}{}^{Fras}(x)$ — описание j-го словосочетания i-го типа; $m(i)$ — количество сочетаний i-го типа, а $n(x)$ — количество типов словосочетаний в словарной статье $V(x)$ с реестровым словом x. В СУЯ всего определено четыре типа словосочетаний: свободные словосочетания (i=1), терминологические словосочетания (i=2), эквиваленты слова (i=3) и собственно фразеологизмы (i=4) [2, 3]. Каждому лексикографическому комплексу — $V^{Lex}(x)$ и $V^{j}{}_{i}{}^{Fras}(x)$ — ставится в соответствие базовая структура:



Здесь в случае $V = V^{Lex}(x)$ в роли Λ_0 выступает заголовочное слово с соответствующими грамматическими характеристиками. Для $V^{j}{}_{i}{}^{Fras}(x)$ Λ_0 представляет словосочетание в реестровой словарной форме плюс определённые грамматические характеристики. Структура правой части идентична как для лексем, так и для словосочетаний любого типа. Стрелками здесь и далее обозначено отношение вложения.

Анализ правых частей $P(x)$ дает следующие структурные элементы: дефиниции лексических значений слова, дефиниции оттенков значений, иллюстрации к каждому значению и каждому оттенку. Обозначим через S_i дефиницию i-го значения реестровой единицы x, $S(Shade)_{ij}$ — дефиницию j-го оттенка i-го значения, $I_i{}^k$ и $I^l{}_{ij}$, соответственно, k-ая иллюстрация *i-го* значения, и l-ая иллюстрация *j-го* оттенка значения. Представим структуру правой части в виде графа:



Через $n(i)$, $m(j)$, $r(k)$ обозначены, соответственно, количество значений, оттенков значений и иллюстраций.

Поскольку между грамматической и лексической семантикой нет абсолютной границы, в лексикографическом представлении лексического значения могут встречаться и грамматические элементы. Для экспликации данного факта в структуре правых частей выделяются две подструктуры: $S(Gram)$, $Shade(Gram)$ — для отображения грамматических эффектов в лексических значениях и $S(Lex)$, $Shade(Lex)$ — для подачи собственно словарных де-

финиций. В структуре иллюстрации различаются подструктуры *I*(*Text*) и *I*(*Passport*) — непосредственно сам текст иллюстрации и библиографическое описание её источника.

Проиллюстрируем структуру словарной статьи на небольшом, но достаточно представительном примере. Из экономии места приводим только иллюстрации для одного значения реестрового слова и только один фразеологизм из блока словосочетаний. (Отметим, что некоторые словарные статьи в СУЯ обладают весьма развитой структурой и насчитывают более тысячи структурных элементов).

Пример 1 (словарная статья с заголовочным словом бувати).

БУВАТИ, áю, áєш, *недок.*

1. Існувати, бути (з відтінком багатократності). ...
2. Відбуватися, траплятися (кілька разів). ...
3. Іноді, час від часу приходити, приїздити кудинебудь, відвідувати когось, щось.
4. Перебувати де-небудь. *Тепер, останніми днями я менше буваю на людях* (М. Коцюбинський); *Коли Марія бувала вдома, посланець чекав, поки вона писала відповідь* (В. Кучер).
5. Уживається у знач. зв'язки в складеному присудку.
6. *тільки* **бува (буває, бувало),** *у знач. вставн. сл.* Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді. ... // Уживається для вираження ймовірності чогось неприємного, недоброго. ... // Уживається для вираження припущення у знач., близькому до може . ... // *у сполуч. із сл.* як. Коли, якщо, випадком. ...

...

◇ ... **(3) Бувати** *(рідко* **бути) / побувати у бувальцях:** а) багато бачити, зазнавати в житті, набувати життєвого досвіду. ... б) *(яких)* опинятися у складних, перев. небезпечних ситуаціях. ... в) мати непривабливий вигляд, довго або часто використовуватися (перев. про одяг); бути не новим. *Вигляд у мене був непривабливий. Сірий простенький костюм, що вже й до цього бував у бувальцях, зовсім зім'явся* (Ю. Збанацький); ...

Левая часть для лексемы L_0(x) ≡ <**БУВАТИ,** áю, áєш, *недок.*>

Структурные элементы правой части приобретают здесь такие значения:

S_1 ≡ <Існувати, бути (з відтінком багатократності)>
S_2 ≡ <Відбуватися, траплятися (кілька разів)>
S_3 ≡ <Іноді, час від часу приходити, приїздити кудинебудь, відвідувати когось, щось>
S_4 ≡ <Перебувати де-небудь>
I_{41}(*Text*) ≡ <*Тепер, останніми днями я менше буваю на людях*>
I_{41}(*Passport*) ≡ <М. Коцюбинський>
I_{42}(*Text*) ≡ <*Коли Марія бувала вдома, посланець чекав, поки вона писала відповідь*>
I_{42}(*Passport*) ≡ <В. Кучер>
S_5 ≡ <Уживається у знач. зв'язки в складеному присудку>
S_6 ≡ <*тільки* **бува (буває, бувало),** *у знач. вставн. сл.* Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді>
S_6(*Gram*) ≡ <*тільки* **бува (буває, бувало),** *у знач. вставн. сл.*>
S_6(*Lex*) ≡ <Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді>
$Shade_{61}$ ≡ <Уживається для вираження ймовірності чогось неприємного, недоброго>
$Shade_{62}$ ≡ <Уживається для вираження припущення у знач., близькому до може>
$Shade_{63}$ ≡ <*у сполуч. із сл.* як. Коли, якщо, випадком>
$Shade(Gram)_{63}$ ≡ <*у сполуч. із сл.* як>
$Shade(Lex)_{63}$ ≡ <Коли, якщо, випадком>

Структурные элементы для словосочетания помечаются верхним индексом *F*. Левая часть словосочетани $L^F_0(x)$ ≡ <**Бува#ти** *(рідко* **бу#ти) / побува#ти у бува#льцях>**. Структурные элементы правой части:

S^F_1 ≡ <багато бачити, зазнавати в житті, набувати життєвого досвіду>
S^F_2 ≡ <*(яких)* опинятися у складних, перев. небезпечних ситуаціях>
$S(Gram)^F_2$ ≡ <*(яких)*>
$S(Lex)^F_2$ ≡ <опинятися у складних, перев. небезпечних ситуаціях>
S^F_3 ≡ <мати непривабливий вигляд, довго або часто використовуватися (перев. про одяг); бути не новим>
$I(Text)^F_{31}$ ≡ < *Вигляд у мене був непривабливий. Сірий простенький костюм, що вже й до цього бував у бувальцях, зовсім зім'явся*>
$I(Passport)^F_{31}$ ≡ <(Ю. Збанацький)>

Все выделенные структурные элементы словарной статьи отображены на структуру компьютерной базы данных, что обеспечивает прямой доступ к каждому из них и возможность построения разнообразных индексных схем. Отметим, что технология работы со словарём в цифровой форме орентирована не на язык разметки, а на структуру лексикографической системы. Такой подход обусловлен чисто прагматическими соображениями. Инструментальный комплекс предназначен для непосредственной работы лексикографов, поэтому целесообразно освободить их от работы по поддержке функционала системы (с этой целью и структура словарной статьи минимальна). Работа над текстом каждой

из базовых структурных единиц ведётся в традиционном режиме, используется минимальный набор средств редактирования (рис. 3). В целом же поддержка структурной целостности словарной статьи и словаря возложена на систему.

В настоящее время система обеспечивает следующие основные функции:

- авторизация и идентификация пользователей;
- добавление и удаление новых пользователей;
- управление правами доступа (просмотр словарных статей, редактирование, доступ к интерфейсам и т. п.);
- добавление новых словарных статей в лексикографическую базу данных;
- удаление словарных статей из базы данных;
- редактирование словарных статей (добавление, удаление структурных элементов в границах заданной структуры словарной статьи, редактирование текста, маркирование проблемных статей);
- динамическое воспроизведение словарных статей в печатном формате или в любом заданном формате визуализации;
- анализ данных (лексикографическая статистика, история лексикографиро-вания каждой словарной статьи с учетом авторизации всех вносимых изменений в базу данных, планирование и учет объемов выполненной работы каждым участником лексикографического процесса, маркирование этапов лексикографической обработки и т. д.);
- выполнение выборок из базы данных по целому ряду параметров (частеречная принадлежность, стилистические и отраслевые ремарки, формулы квазисемантики и т. д.);
- создание SQL-запросов и формирование подсистем по заданным характеристикам.

На рисунках 1–3 продемонстрированы некоторые окна пользовательских интерфейсов системы.

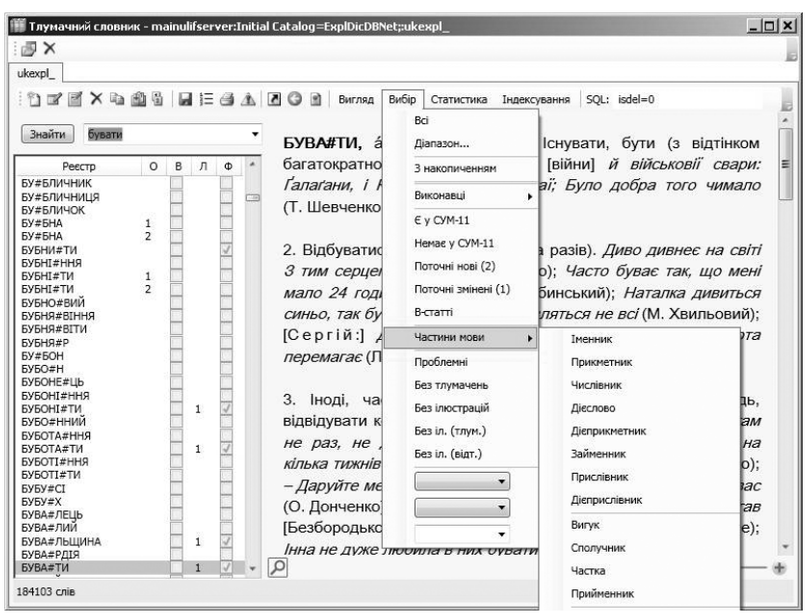


Рис. 1. Окно главного пользовательского интерфейса

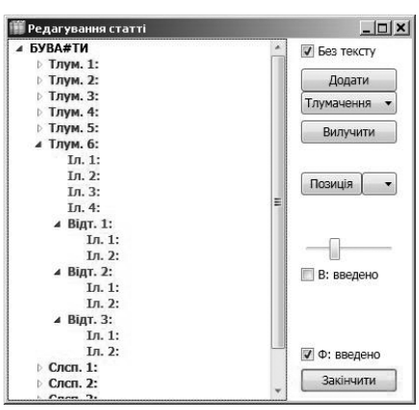


Рис. 2. Окно навигации по структуре словарной статьи

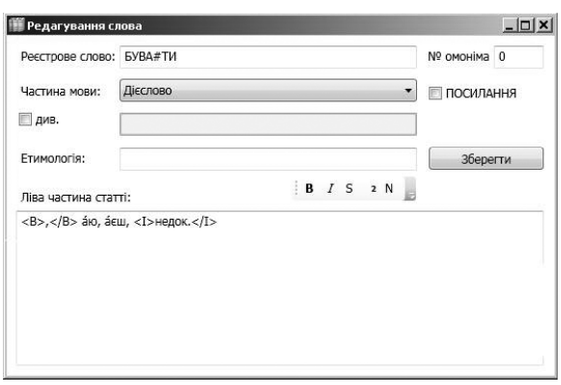


Рис. 3. Окно редактирования левой части словарной статьи

Виртуальная лексикографическая лаборатория создана в среде разработки Microsoft Visual C# 2008 Professional Edition. Она работает в операционных системах Microsoft Windows XP/2003, Vista или Windows 7 под управлением Microsoft.NET Framework версии 3.5 SP1. Комплекс имеет многоуровневую архитектуру: сервер базы данных отвечает за связь с лексикографической базой данных, функции получения и сохранения данных; сессионный сервер устанавливает сеансы работы для отдельных пользователей, управляет привилегиями и устанавливает уровни доступа; клиентская программа предоставляет интерфейс пользователя. Таким образом, программный комплекс ориентирован на работу в сети — как локальной, так и глобальной, поскольку использование технологии создания распределенных сервис-ориентированных систем Windows Communication Foundation (WCF) для взаимодействия между отдельными уровнями комплекса позволяет ему эффективно функционировать в среде Интернет. В настоящий момент решается задача перевода комплекса на технологию работы с объектно-реляционной моделью данных Entity Framework 4.0, которая обеспечивает большую гибкость работы и независимость программных интерфейсов от физической реализации базы данных. В настоящее время используется технология ADO NET для работы с данными (СУБД Microsoft SQL Server 2008).

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.

Хотя работа над текстом словаря будет продолжаться ещё в течение нескольких лет (вообще, предполагается, что он будет развиваться постоянно, отслеживая языковую динамику украинского языка), как целостный цифровой лексикографический продукт уже в настоящее время Словарь готов для использования при заполнении лакун национальной лексикографии: это семантический словарь и представительные (с реестром 200–300 тысяч единиц) украинско-иноязычные словари.

Литература

1. *Русанівський В. М., Широков В. А.* Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії // Мовознавство., 2002, № 6. С. 7–48.
2. *Словник української мови:* В 11 томах. К.: Наукова думка, 1970–1980.
3. *Широков В. А.* Елементи лексикографії. К.: Довіра, 2005. 304 с.
4. *Oxford English Dictionary.* [Электронный ресурс] (http://dictionary.oed.com/).