Электронные словари Globus Software

И. Н. Ларченков

Анализируется опыт создания электронных словарей департаментом прикладной лингвистики компании “Globus Software”. Раскрываются принципы, положенные в основу системы электронных словарей компании, их связь с классическими лексикографическими нормами и дополнения, которые, по мнению авторов проекта, делают электронный словарь электронным.
Обсуждается состав словарной статьи, а также язык ее разметки (DML). Используя открытый формат DML, авторы словарей имеют возможность обмениваться данными с другими словарями или интегрировать данные в различные системы публикации. Рассматриваются алгоритмы управления словарной статьей, использующие специализированный модуль пополнения и модификации записей словаря. Модуль позволяет создавать словари большой сложности, не вдаваясь в подробности структуры языка разметки.
Поясняются функциональные особенности использования интегрированных в систему грамматических словарей и уникальный алгоритм их пополнения. При пополнении грамматического словаря алгоритм позволяет за минимальное количество шагов определять все формы русского или английского слова.
Описываются функции и способы, позволяющие авторам публиковать свои словари в классическом и в электронном виде. Термином “электронный словарь” в

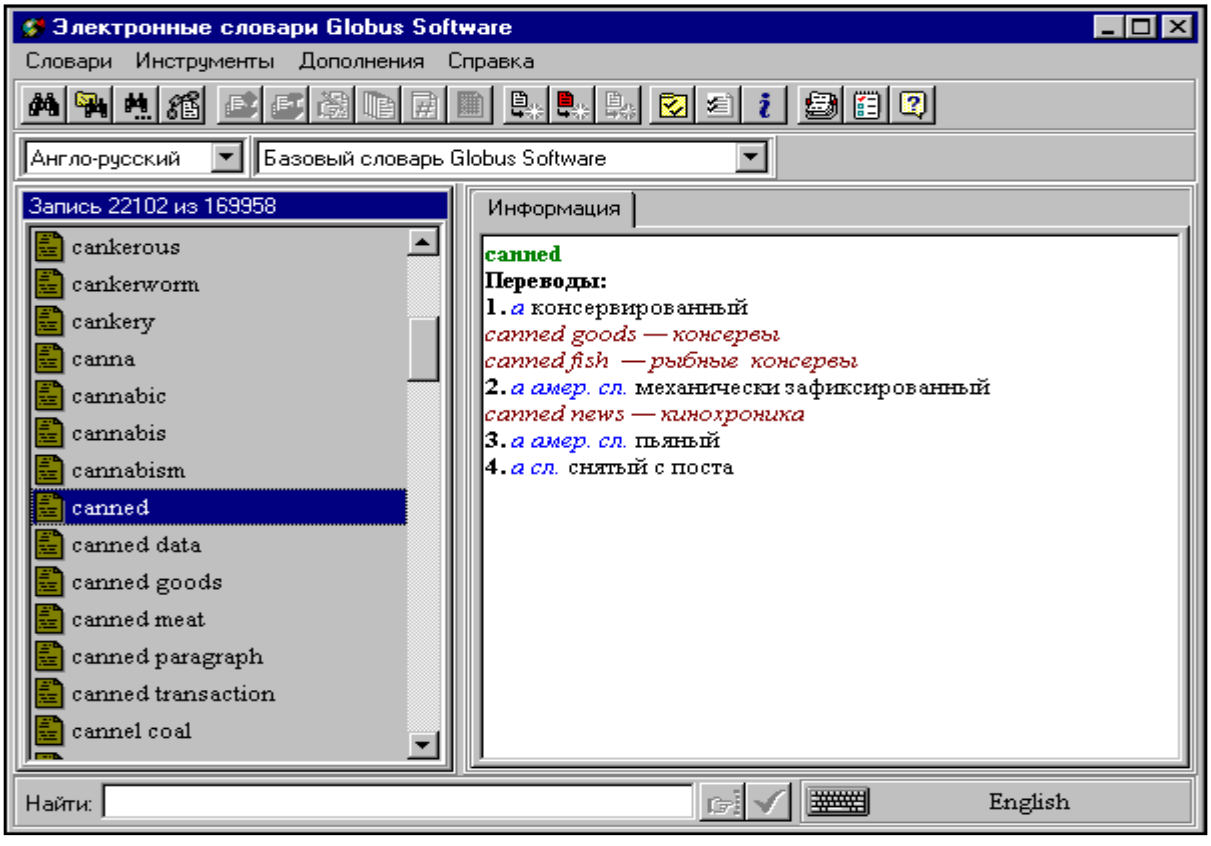

Рисунок 1 Главное окно электронных словарей Globus Software

настоящее время уже трудно кого-либо удивить. При всем при этом единого мнения на то, что следует считать электронным словарем, нет, не было и, на наш взгляд, быть не может. Как и сами словари, смысл, вкладываемый в это понятие, является авторским и отражает то видение вопроса, которым располагает лексикограф, его опыт, знания и даже профессиональную принадлежность.

Компьютерная лексикография, несмотря на бурное развитие вычислительных средств и средств коммуникации, также, на наш взгляд, не принесла что-либо кардинально изменяющее основные лексикографические принципы. Основополагающими по этой теме до сих пор являются труды Ю. Д. Апресяна, А. И. Смирницкого и ряда других ученых. При этом все, что было верно в традиционной лексикографии, тем более является верным в ее электронной версии.

К великому сожалению, надо признать, что использование компьютерной техники не слишком изменило и экономическую составляющую вопроса: создать хороший словарь было и есть дорого и долго. Именно поэтому практически все электронные словари до сих пор в значительной степени базируются на широко известных традиционных классических словарях.

Разработчики компании “Globus Software”, (группа компаний “Светон”) провели анализ существующих “бумажных” и популярных электронных словарей и выработали свою, в какой-то степени оригинальную, но близкую к традиционной, концепцию построения подобных систем.

Мы решили не противопоставлять традиционную лексикографию ее компьютерному “электронному” эквиваленту. Считая, что традиционная лексикография является наиболее проработанной и полной, она была взята за основу при построении комплекса одноименных (Globus Software) словарей. Таким образом, основной идеей при создании наших словарей было электронным способом реализовать классические лексикографические принципы и подходы. Великолепно проработанная теория и практика создания классических словарей должна дополняться широкими функциональными возможностями, которые предоставляет электронный носитель.

Компактность хранения информации в электронном виде позволила существенно расширить поддерживаемую словарную статью, оставив ее в пределах старых добрых традиционных норм. Вернее, если быть точным, мы объединили в единую словарную статью такую информацию как:

- Систему признаков
- Фонетическую информацию
- Переводы и комментарии к ним
- Список связанных с ключом записей (список устойчивых сочетаний, типовых фраз и т.д.)
- Текстовые комментарии
- Толкование записи
- Синонимические ряды
- Антонимические ряды
- Данные, определенные пользователем

Все эти поля не являются обязательными, и лексикограф имеет возможность отключить любые из них.

Если ни одно из перечисленных интегрированных полей не подходит для хранения той информации, которая требуется по авторскому замыслу, поддерживается пользовательский тип данных, в который можно внести любую текстовую информацию.

Сложная система единой словарной статьи привела к необходимости создать специализированный формат ее текстового описания. В компании разработан язык описания словарной статьи DML, который позволяет осуществлять ее разметку в текстовом виде. По

своей структуре DML напоминает широко используемый язык HTML. Он является теговым, при этом каждый тег построен по принципу, позволяющему компилятору игнорировать все теги, которые им не поддерживаются. Это в какой-то степени дает возможность реализовать совместимость компиляторов не только снизу вверх (от младшей версии к старшей), но и сверху вниз. Компилятор с языка DML встроен непосредственно в интегрированную оболочку и является составной частью всех версий наших словарей.
Сразу оговоримся, мы не предлагаем разработанный нами язык в качестве какого-либо де-юре или де-факто стандарта. Это язык, используемый в комплексе словарей Globus Software, и не более того. Он позволяет модифицировать словарную статью и, при желании, конвертировать собственные словари в формат, поддерживаемый другими электронными словарями или издательскими системами. DML является открытым, общедоступным и совершенно бесплатным. Ниже приведен пример разметки словарной статьи в формате DML.

```
<ENTRY>
<KEY> afternoon </KEY>
<DOMAIN> сущ. </DOMAIN>
<TRANSLATEDATA>
<TTRANSLATE> время после полудня </TTRANSLATE>
<TSAMPLE>
I always have a rest in the afternoon. Я всегда отдыхаю днем.
</TSAMPLE>
</TRANSLATEDATA>
<SEEALSO>
<SKEY> good afternoon! </SKEY>
<STRANSLATE> добрый день! </STRANSLATE>
</SEEALSO>
<SEEALSO>
<SKEY> in the late afternoon </SKEY>
<STRANSLATE> к вечеру </STRANSLATE>
</SEEALSO>
<SEEALSO>
<SKEY> this afternoon </SKEY>
<STRANSLATE> сегодня днем </STRANSLATE>
</SEEALSO>
</ENTRY>
```

Однако далеко не всегда удобно использовать словарную статью в текстовом представлении. Электронные словари Globus Software снабжены специализированным редактором словарной статьи, который предоставляет доступ ко всем данным в диалоговом режиме. Так же, как и компилятор с языка DML, редактор встроен непосредственно в интегрированную оболочку и является составной частью всех версий наших словарей. Безусловно, любая модификация данных словаря возможна только при наличии авторских прав (защищенных паролем) на словарь в целом.

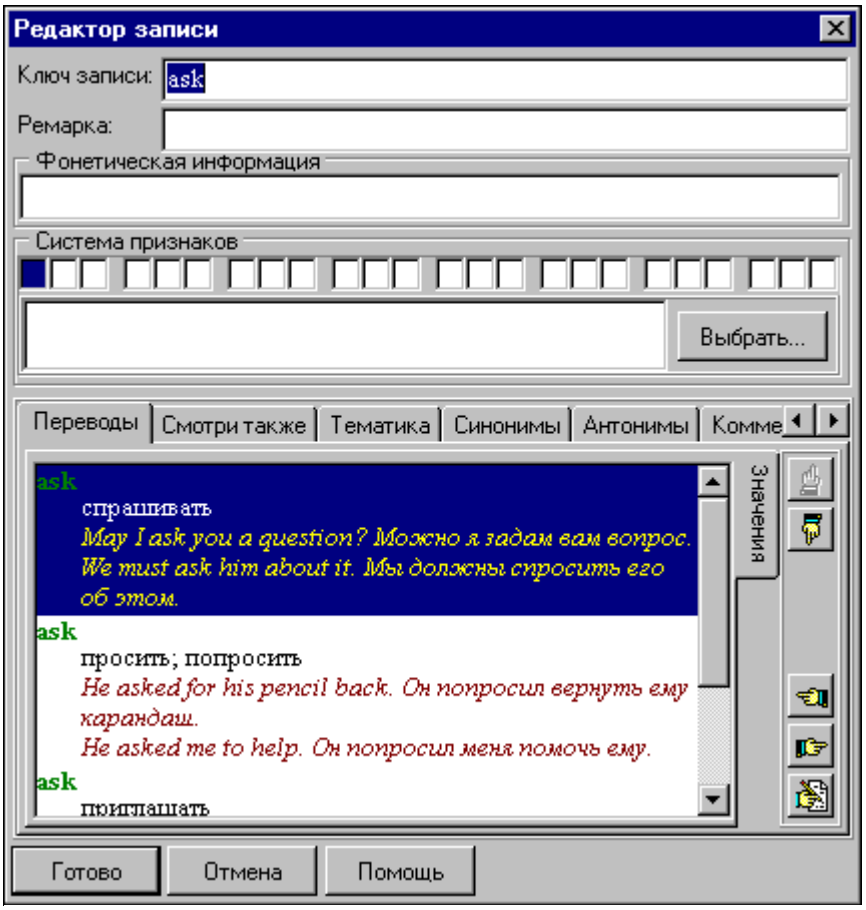

Рисунок 2 Окно редактора словарной статьи

Диалоговое окно редактора ди-намично настраивается и ото-бражает только те поля, кото-рые присутствуют в редакти-руемом словаре. Автор словаря имеет право заблокировать дос-туп к любому полю словарной статьи, что позволяет гибко настраивать редактор под текущие решаемые задачи. Редактор словарной статьи имеет ряд особенностей, кото-рые делают процесс заполне-ния словарной статьи данными максимально быстрым и удоб-ным. Например, при ведении англо-русского словаря, если пользователем не заблокирова-на такая возможность, будет происходить автоматическая смена расклада клавиатуры на язык, на котором, как правило, заполняются данные этого поля. Каждой словарной статье мож-но приписать ряд символьных признаков, которые могут ис-пользоваться при поиске. В этом случае система позволит вести библиотеку типовых наборов признаков, что тоже существенно помогает при создании словаря.

Таким образом, интегрирован-ный в систему словарей Globus Software редактор данных является мощным средством модификации и пополнения словаря. При этом лексикограф может вообще ничего не знать о языке DML и его тегах.

Электронные словари Globus Software имеют гибкую функцию настройки и управления словарями. Пользователь может исключить словарь из системы или добавить новый. Изменяя последовательность словарей в списке, устанавливаются приоритеты каждого словаря по отношению к другим словарям.

При наличии авторских прав имеется возможность изменить свойства словаря Globus Software. Словарь можно переименовать, ввести или изменить авторский комментарий к словарю, изменить номер версии или сборки, изменить состав словарной статьи, список используемых шрифтов, пароль администратора (автора), установить или снять флаг разрешения модификации словаря или прав на его публикацию. Автор может записать информацию о себе в виде фотографии и комментария, которые будут храниться непосредственно в файле словаря и отображаться в окне с информацией о словаре, доступном всем пользователям.

Отметим такое важное свойство нашей системы как публикация созданных в ней словарей. Как было отмечено, за основу построения комплекса были приняты те нормы и правила, которые практикуются при создании традиционных “бумажных” словарей. Классическая

лексикография была просто переведена на “электронный” носитель. Мы считали и считаем, что традиционный словарь еще долгое время был, есть и будет вершиной труда лексикографа. При этом подходе было бы логичным, после создания электронного словаря, выполнить “обратную” операцию: создать макет традиционного бумажного словаря. Возможность такого экспорта данных реализована, начиная с версии 2.0. При наличии соответствующих прав можно автоматически создавать прототип оригинал-макета словаря в виде документа формата RTF или HTML. Этими же средствами реализован экспорт словаря в текстовый файл формата DML, что позволяет объединять несколько словарей или конвертировать данные в формат других электронных словарей.

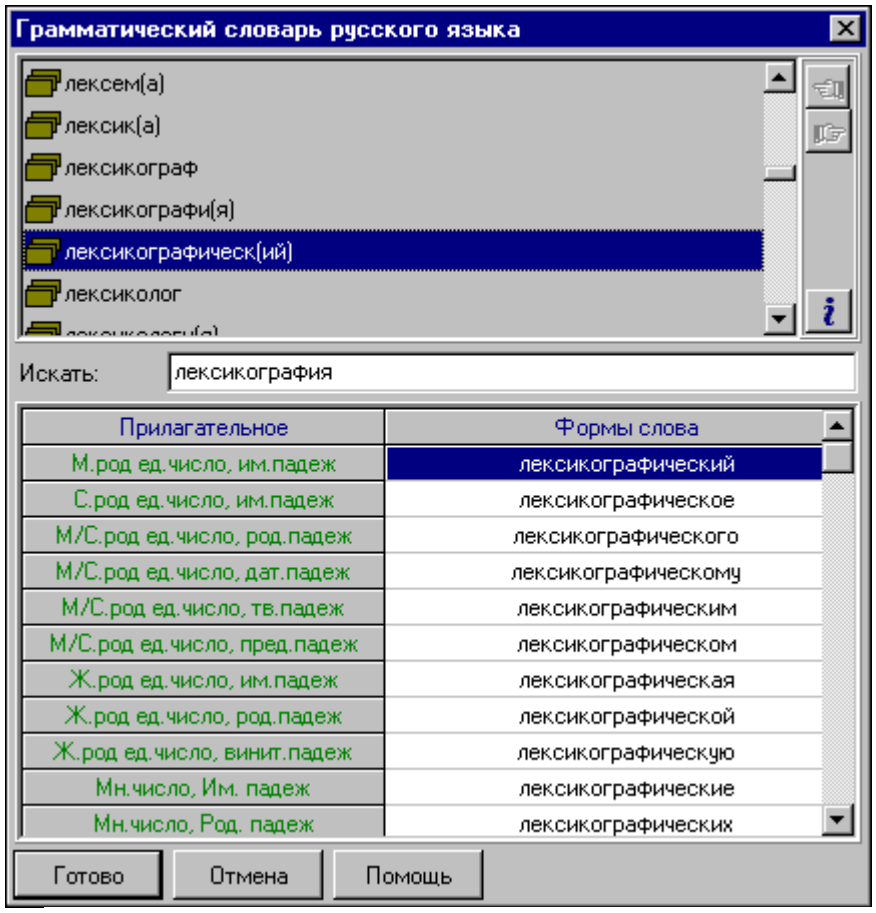

Рисунок 3 Вид диалогового окна морфологической библиотеки русского языка

Комплекс Globus Software располагает мощными биб-лиотеками морфологичес-кого анализа русского и ан-глийского языков. Модуль морфологического анализа русского языка базируется на классическом граммати-ческом словаре русского языка А. А. Зализняка, и содержит в себе более 100 000 входов. В нем реали-зована уникальная функция пополнения электронного грамматического словаря. Уникальность этой функ-ции в том, что она работает с использованием ориги-нального алгоритма мини-мизации действий. Как по-казал анализ, при исполь-зовании этого алгоритма пользователю требуется выбрать не более 4-х форм русского слова из постоян-но сокращающегося списка образцов. При этом алго-ритм автоматически до-страивает всю оставшуюся парадигму русского слова.
Модуль английской морфо-логии создан на базе соб-ственного словаря компа-нии объемом свыше 50000 записей. В этом модуле также работает упрощенный алгоритм минимизации действий.
Все модули морфологического анализа имеют дружественный интерфейс, позволяющий по любой форме слова просмотреть всю его парадигму. Морфологический анализ используется при поиске слов и выражений в наборе словарей Globus Software.
Стало доброй традицией делать электронные словари “звучащими”. Мы не стали изменять ей, однако вынесли озвучивание слов и выражений в отдельное приложение (вернее, несколько приложений). Такой подход мы считаем более чем оправданным. Как правило,

фонетическая поддержка подобного рода требуется при обучении языку, и тот, кому она необходимо, желает получить дополнительно ряд других возможностей по тренингу своего словарного запаса. Таких потенциально необходимых возможностей оказалось столь много, что мы посчитали разумным вынести эту функцию за скобки основной программы. Так в составе комплекса Globus Software появились модули тренинга лексического запаса.
Кроме группировки слов по темам и использования звукового ряда, в нем реализованы такие возможности:

- Система “карточного” пополнения словарного запаса, которая хорошо зарекомендовала себя при обучении языкам.
- Создание урока из нескольких тематических групп и подгрупп
- Система проверки знаний
- Занесение трудных для запоминания слов в отдельный словарь и использование его как базового

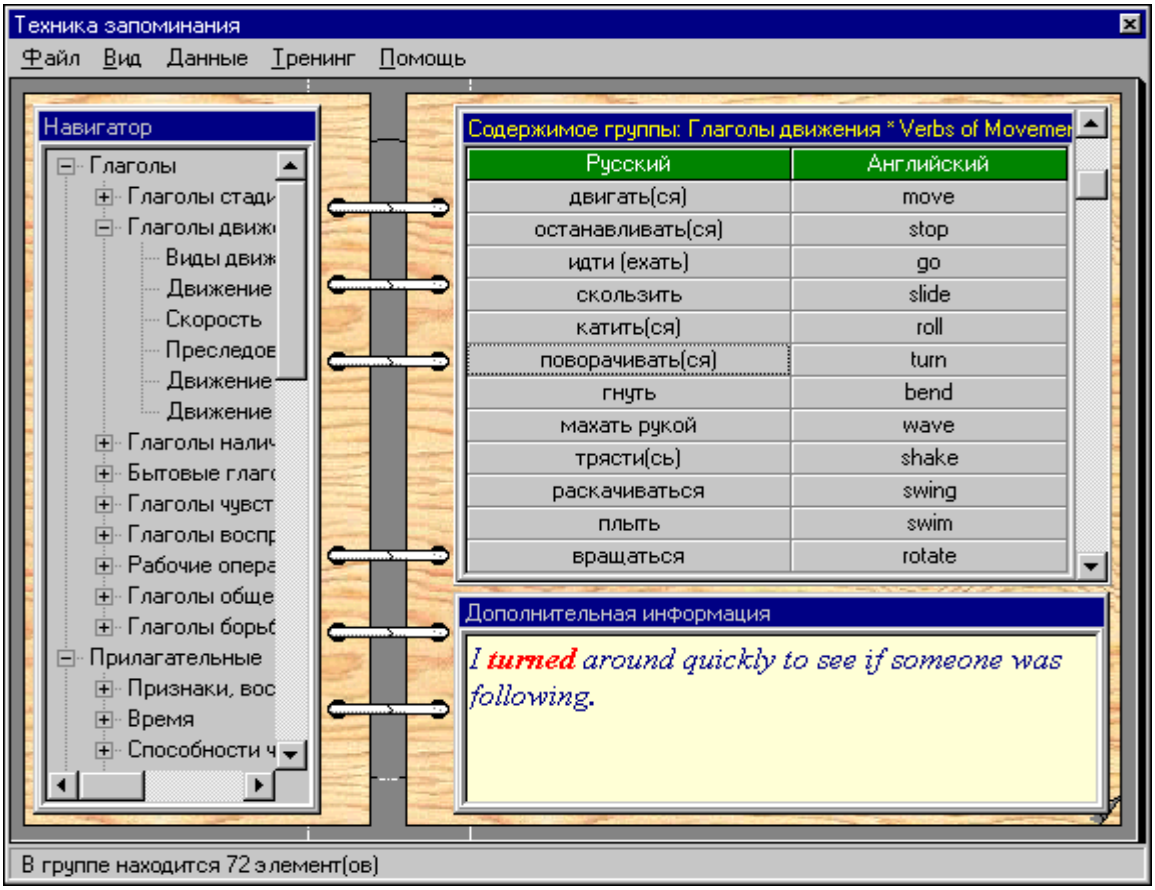

Рисунок 4 Главное окно модуля тренинга словарного запаса

Электронная версия традиционной лексикографии еще имеет много не реализованных методов и приемов. Планы компании на следующую версию словарей включают в себя как значительное увеличение качества и количества базовых словарей, входящих в систему, так и существенное увеличение сервисных функций. Впрочем, это дело будущего. Однако уже сейчас авторы комплекса уверены в том, что подходы и принципы, положенные в основу текущей версии, позволят Globus Software занять достойное место на рынке лингвистического программного обеспечения.