Комплекс по разработке индивидуальных и/или корпоративных электронных толковых словарей

Хахалин Г.К., Богданов Н.К., Платонов С.В.
khakhalin@got.mmtel.ru

Аннотация

Комплекс по созданию толковых словарей обеспечивает разработку и поддержку электронного словаря пользователя с заполнением на любую выбранную предметную область. Комплекс состоит из оболочки электронного словаря с доступом к словарным статьям по любой словоформе и системы наполнения на конкретную предметную область. Он служит пользователю или группе пользователей средством для создания индивидуальных и/или корпоративных словарей. Электронные толковые словари, реализованные с помощью данного комплекса, могут быть использованы в процессе принятия решений или в процессе обучения. Словарь обеспечивает возможность адаптации электронной оболочки к любому словарю толкового типа за счет пополнения декларативных компонентов словаря. Разнообразие режимов работы со словарем позволяет: вводить любое слово в произвольной форме и получить его толкование, не задумываясь о стандартной форме; получать всю предысторию работы со словарем, возвращаясь на любой из предыдущих этапов; работать со словарной толковой статьей как с обычным текстовым файлом и т.д. Известные электронные толковые словари “зашиты” и не предоставляют возможность пользователю создать “собственный” словарь. Другие электронные словари, появляющиеся на рынке программных продуктов, являются не толковыми, а двуязычными и разработаны по несколько иной технологии.

Ключевые слова: лингвистические словари, электронный толковый словарь, лингвистическая база данных, морфологический анализ, нормализация словоформы, настройка на предметную область, средство обучения, индивидуальный и/или корпоративный электронный словарь

Введение

В языкознании нет достаточно строгой дифференциации словарей [Маслов, 1997]. Принято различать лингвистические и нелингвистические словари. Некоторые словари при этом носят промежуточный характер. В лингвистических словарях обычно описывают лексические единицы языка (слова и фразеологизмы). В нелингвистических словарях лексические единицы служат лишь опорными точками для сообщения тех или иных сведений о предметах и явлениях внеязыковой деятельности. К лингвистическим словарям относятся: толковые, переводные, частотные, грамматические и др. словари. К специальным лингвистическим словарям относятся: фразеологические словари (переводные и одноязычные), словари "крылатых слов", словари народных пословиц и поговорок, словари синонимов, антонимов, омонимов и др. словари. Переходными от лингвистических к нелингвистическим словарям являются словари терминов различных наук и отраслей техники. Выделяются также универсальные словари: одновременно толковые и энциклопедические, включающие этимологические и исторические справки и снабженные в нужных случаях рисунками. Примерами последних словарей являются: "словари Лярусса" (по имени фр. издателя) и английские "словари Вебстера" (по имени первого составителя этих словарей).

В области электронных словарей также существует многообразие различных их вариантов: от простого перевода книжного словаря в электронный вид до разработок, в которых учитываются особенности электронной "подачи" информации и работа на компьютере. Ниша электронных словарей всех видов в настоящее время интенсивно заполняется. На рынке программных продуктов существуют двуязычные электронные словари, служащие для целей

автоматизации процесса перевода иноязычных текстов (например, система Lingvo или коллекция электронных словарей Мультилекс). Есть разработки и общезначимых электронных словарей (например, "Русский филолог"[Лахути и др., 1993], который необходим как справочник для лингвиста и "обилен" для широкого пользователя, поскольку содержит достаточно много информации о русском языке: о толковании, словообразовании, управлении, синонимии и грамматике). Есть специализированные электронные словари, которые предназначены для узкого круга специалистов (например, электронный словарь языка Грибоедова А.С., который задуман как полноценный текстологический справочник для лингвиста [Поляков, 1999]).

Двуязычные электронные словари редко объясняют термины. Существующие электронные толковые словари "зашиты", т.е. они не позволяют самому пользователю создать свой словарь или модифицировать уже существующий.

Целью данной разработки является предоставить пользователю или группе пользователей возможность и средства, позволяющие самим создавать электронный толковый словарь или словарь терминов с его заполнением на любую предметную область для индивидуального и/или корпоративного использования. Применять такой словарь можно как аналог "книжного" толкового или терминологического словаря, так и как средство для принятия решений и обучения в различных проблемных областях, где требуется разъяснение терминологии и определение понятий. Первоначальный вариант комплекса для ДОС'а был представлен как "побочный" продукт лингвистического транслятора в [СЭТС, 1996; Хахалин, 1998]. Данная версия разрабатывается для ОС Windows 9x/NT. Естественно, некоторые технические решения претерпели существенные изменения в связи с выбором другой операционной среды.

1. Функции комплекса

Книжный словарь толкового типа дает толкование значений слов какого-либо языка средствами этого же языка. Толкование дается с помощью логического определения концептуального значения, посредством подбора синонимов или в форме указания на грамматическое отношение к другому слову. В некоторых толковых словарях значения слов раскрываются с помощью рисунков. Эмоциональные, экспрессивные и стилистические коннотации указываются посредством специальных помет (неодобр., презр., шутл., ирон., книж., разг. и т.п.). Отдельные значения иллюстрируются примерами: типичными сочетаниями или же цитатами. Обычно толковые словари дают также грамматическую характеристику, указывая с помощью специальных помет на часть речи, грамматический род существительного, вид глагола и т.д. и приводя в нужных случаях кроме словарной и некоторые другие формы данного слова. В той или иной форме указывается и произношение слова (например, в русских толковых словарях ударение). Лексикографическая структура словаря включает следующие составляющие: словарная статья, заголовочное слово и словник. Словарная статья - это абзац или несколько абзацев словаря, дающих информацию, относящуюся к одной лексической единице (иногда к нескольким взаимосвязанным единицам). Статья начинается заголовочным словом (иногда сочетанием), обычно выделенным особым шрифтом. Совокупность всех слов, рассматриваемых в словаре, называется словником этого словаря.

Работа человека с книжным толковым словарем достаточно известна. Книжный словарь предоставляет человеку возможность при чтении текстов по слову найти его толкование; "проштудировать" толковую статью; также, встречая в словарной статье "интересные" слова, обращаться к их толкованиям; перелистывать страницы словаря, обучаясь новой терминологии и связывая ее с уже известными знаниями, и т.д.

Комплекс должен обеспечивать все эти функции и реализовывать их так, чтобы для человека они выглядели естественным образом. Он должен быть простым в обращении и содержать только ту информацию и в том объеме, которые необходимы пользователю с учетом направленности словаря на принятие решений и обучение. Кроме этого комплекс должен обеспечивать поддержку электронного словаря с заполнением на любую выбранную пользователем предметную область. С учетом машинной реализации должно быть обеспечено разнообразие режимов работы со словарем, что позволило бы пользователю: набирать или, просматривая текстовый файл, отмечать любое слово в произвольной форме и получать его толкование, не задумываясь о стандартной форме слова; получать предысторию работы со словарем с возможностью возвращения на любой из предыдущих этапов; работать со словарной толковой статьей как с обычным текстовым файлом; получать для омонимичных словоформ список лексем-омонимов и т.д.

Следовательно, в рамках комплекса должны быть реализованы следующие функции:

- нормализация входной словоформы (получение лексемы) с помощью функций морфологического анализатора;
- вывод списка лексем в случае омонимии слова (режим "Альтернативы");
- поиск данных (лексем и толковых словарных статей) непосредственно по введенной словоформе;
- вывод найденных данных (толковой словарной статьи или списка лексем) или сообщение об их отсутствии;
- специальное форматирование словарной статьи (выделение ударения, примеров и т.п.);
- поиск по слову, встречающемуся в самой толковой словарной статье;
- просмотр вводимых в сеансе работы слов (режим "Предыстория");
- ввод толкования слова с проверкой на уникальность и с возможностями редактирования вводимых данных;
- вывод списка "похожих" слов на входное слово с ошибкой или на ввод отсутствующего в словнике слова (режим "Варианты");
- интеграция с текстовым процессором типа Microsoft Word.

Конечный продукт должен быть выполнен в виде самостоятельного пакета, запускаемого с помощью исполняемого модуля. Система должна функционировать в ОС Windows 9x/NT. Инсталляционный пакет должен включать все необходимые компоненты и не требовать дополнительной настройки стороннего программного обеспечения или устройств. Большинство перечисленных функций реализованы в данной версии 1.0 для Windows 9x/NT.

2. Структура комплекса

Комплекс состоит из оболочки электронного толкового словаря и из средств наполнения (добавления) толковых словарных статей.

Оболочка словаря включает в себя лингвистическую базу данных и морфологический анализатор (для русского языка). Оболочка реализует нормализацию введенной или выбранной словоформы с помощью морфологического анализатора, поиск толковой словарной статьи по нормализованной словоформе (лексеме) и отображение в соответствующем окне найденной статьи.

Адаптация (настройка) словаря к выбранной предметной области осуществляется средствами наполнения. Они позволяют пользователю динамически добавлять в словарь толковые словарные статьи по мере необходимости, форматируя тексты статей с помощью специализированного редактора. После добавления статьи ее поиск и выборка могут быть произведены уже в текущем сеансе.

2.1. Лингвистическая база данных

Лингвистическая база данных функционально содержит две базы: заполненную базу морфологических словарных статей и "пустую" базу толковых словарных статей. В данной версии словаря база морфологических статей недоступна для пользователя и может быть модифицирована (сокращена или пополнена) только разработчиком. Она задана в объеме словаря Зализняка А.А. (более 130 тысяч словарных входов) [Зализняк, 1980]. База толковых статей заполняется разработчиком или самим пользователем с помощью средств пополнения словаря. Рассмотрим представление лингвистических данных.

Формат хранения данных

Оптимальным форматом хранения множества однотипных данных являются таблицы реляционной базы данных (БД). Данный подход обеспечивает простоту и качество хранения, обработки и записи данных. Данные в таблицах хранятся в структурированном и компактном виде. Благодаря возможности использования непроцедурного языка структурированных запросов (SQL) нет необходимости заботиться о физическом формате данных и создавать процедуры чтения и записи из файла, учитывающие формат хранения данных. Система управления базами данных (СУБД) обеспечивает автоматическое преобразование непроцедурных команд на обработку и преобразование данных в набор команд обработки физических данных (таких как открытие файла, поиск в файле, выделение блоков символов, их чтение или запись и преобразование к формату записей). Помимо указанных преимуществ, СУБД автоматически контролирует целостность данных, их соответствие заявленному типу, соблюдению ограничений (например, на уникальность), кэширование данных и т.п.

Автоматическое кэширование средствами СУБД является дополнительным мощным средством оптимизации работы прикладных программ. При кэшировании, выбираемые данные помещаются в оперативную память и при повторном запросе (что при работе со словами

является частой операцией) просто считываются из нее, что значительно ускоряет работу словаря. Требования к различным выделениям в рамках толковой словарной статьи не накладывают ограничений на использование СУБД, так как в поля таблиц могут быть записаны любые символы, а приложение может трактовать некоторые из них как спецсимволы и соответствующим образом преобразовывать текст. Дополнительными плюсами использования СУБД для хранения данных и работы с ними, является открытость формата, переносимость и универсальность. Создаваемые блоки данных, при учете удовлетворения основным требованиям (имя таблицы, названия и типы полей) могут легко добавляться к существующим блокам или при необходимости замещать их.

Следует отметить, что все манипуляции с данными осуществляются в виде работы с записями таблицы, т.е. запись данных осуществляется блоками с именованными элементами («записи» или строки), а выбранные данные представляют собой табличный блок из нескольких строчных блоков с именованными элементами (некоторая таблица). При этом элементы, имеющие одинаковый номер в строках имеют одинаковое имя и тип в таблице (поле или столбец). Получается своеобразный двумерный массив с именованными столбцами различных типов и нумерованными строками, включающими набор данных, соответствующий набору столбцов. О данном блоке может быть легко получена необходимая информация (количество строк, номер текущей строки, количество столбцов и их типы). Такими блоками легко манипулировать: осуществлять различные переходы по строкам с возможностью циклов, обращение к элементам строки по имени поля или по номеру, преобразование данных, вторичную выборку и отображение всего содержимого блока.

Использование СУБД не накладывает значительных ограничений на скорость работы с данными по сравнению с файловой системой. Появление временных задержек может возникать вследствие использования не оптимальных путей работы с данными, попыток максимально универсализировать работу с данными и в связи с использованием объектно-ориентированной среды программирования и современной ОС. Но с учетом возросшей вычислительной мощности и скорости современных ЭВМ, этот недостаток будет иметь малое значение.

Структура хранения лингвистической информации

Для представления морфологических данных используется морфологическая модель, описанная в [Мальковский, 1985]. В модели множество слов разбивается на два основных класса: неизменяемые (Н-слова) и изменяемые (И-слова). Конкретная словоформа характеризуется основой и флексией, за которой может следовать постфикс. С основой И-слова, Н-словом, флексией и словоформой связывается описание грамматических характеристик, которые определяют сочетаемость основ и флексий, а также синтаксические признаки.

Грамматические характеристики включают: морфологический (словоизменительный) класс (М-класс), парадигматический класс (П-класс), чередование и исключение. Синтаксическим показателем является синтаксический класс (С-класс). Синтаксические и словоизменительные признаки определяются набором значений грамматических переменных (ГП): одушевленность, род, число, залог и т.д. Понятие М-класса является уточнением традиционного понятия "часть речи". Особенности сочетания основ с флексиями выделены в исключения нескольких типов. Несовпадение номеров М-класса и С-класса относится к типу МС. Исключение сочетания основы с "нестандартной" флексией относится к типу МИ (*мастера*, стандартный вариант *артисты*). К исключениям также относятся: наличие супплетивных форм (форма *лучше* у слова *хороший*) и "ложная" возвратность при наличии у слова частицы "ся" (*выдающийся*). К особенностям словоизменения относятся и чередования в основе (подробнее см. [Мальковский, 1985, с. 101-110].

Морфологическая информация и толковые словарные статьи хранятся в табличном виде. Для морфологических данных используются три таблицы: таблица основ, таблица лексем и таблица морфологических признаков (см. таблица 2.1, 2.2 и 2.3 соответственно). Такое разделение информации на три таблицы выполнено в целях ускорения работы системы. Принятая структура данных позволяет хранить до 1 млн. основ и лексем (максимальной длины в 24 символа).

Таблица 2.1. ***voc_morf_osn*** – словарь основ

№	Атрибут	Тип данных	Описание
1	code	numeric[6]	Уникальный ключ
2	osnova	character[24]	Основа

Таблица 2.2. ***voc_morf_lex*** – словарь лексем

№	Атрибут	Тип данных	Описание
1	code	numeric[6]	Ключ
2	lexema	character[24]	Лексема

Таблица 2.3. ***voc_morf_inf*** – словарь морфологических признаков

№	Атрибут	Тип данных	Описание
1	code	numeric[6]	Ключ
2	izm_flag	character[1]	Флаг изменяемости {S|R|W}
3	pclass_num	numeric[2]	Номер П-класса
4	pclass_ch	character[48]	Список флексий, при которых происходит чередование
5	mclass_num	numeric[1]	М-класс
6	gp_osnova	character[8]	Значение ГП
7	ms_iskl	character[10]	МС-исключение
8	mi_iskl	character[10]	МИ-исключение
9	suppl_ots	character[10]	Супплетивная форма
10	vozvr_iskl	character[10]	"Ложная" возвратность

Для представления толковой статьи используется четвертая таблица, в которой заданы лексема и текст словарной статьи (см. таблицу 2.4). Описание толковой статьи в memo-поле может занимать до 64 Кбайт (т.е. порядка 60 тыс. символов без учета форматирующих HTML-тэгов). Хотя размер и не может быть формально увеличен, использование гиперссылок и возможность встраивания JavaScript-кода в описание позволяет практически неограниченно расширять объем словарной статьи, в том числе вставлять изображения, схемы, видеофайлы и т.п.

Таблица 2.4. ***slov*** – описания толковых словарных статей

№	Атрибут	Тип данных	Описание
1	word	character[24]	Определяющее слово
2	topic	memo	Словарная статья (с включенными тэгами HTML-форматирования)

Например, для слова «арфа» содержание записей в таблицах будет следующим:

В таблице ***voc_morf_osn***:

osnova	**code**
арф	8564

В таблице ***voc_morf_lex***:

code	**lexema**
8564	арфа

В таблице ***voc_morf_inf***:

code	**izm_ flag**	**pclass_ num**	**pclass_ ch**	**mclass_ num**	**gp_ osnova**	**ms_ iskl**	**mi_ iskl**	**suppl_ ots**	**vozvr _islk**
8564	S	11		7	2200				

В таблице ***slov***:

word	**topic**
арфа	<FONT COLOR="RED">а</FONT>рфа Щипковый музыкальный инструмент в виде большой треугольной рамы. <I>Играть на арфе.</I>

В поле текста толковой статьи (столбец topic) заданы форматирующие HTML-тэги для выделения цветом ударения, для продолжения текста с новой строки и для выделения курсивом примера использования слова "арфа".

2.2. Морфологический анализатор

В качестве метода морфологического анализа взят за основу метод, предложенный в [Мальковский, 1985]. На вход анализатора поступает словоформа. Сначала анализатор пытается воспринять эту словоформу как Н-слово. Если поиск неуспешен, делается попытка отделить постфикс, а затем строятся все возможные (с учетом длины словоформы) ее разбиения на основу и флексии, последовательно рассматривая словоформу как основу с пустой флексией и с флексиями 3, 2 и 1. Расщепленные части ищутся в таблице основ и в соответствующих полях таблицы морфологических признаков. В случае удачного поиска анализ прекращается и в результате на выход анализатора подается найденная по основе лексема. Например, для входной словоформы *арфу* анализатор выдаст лексему *арфа*. Если в морфологическом словаре явно указана возможность неоднозначного членения, анализатор проверяет все возможные интерпретации и выдает на выходе список омонимов для заданной словоформы. Например, по входной словоформе *цари* будут выданы две лексемы: существительное *царь* и глагол *царить*. Если проверка правильности членения не дала положительного результата или если при правильном членении не было найдено в словаре соответствующей основы, то анализатор входную словоформу воспринимает как форму незнакомого слова и выдает соответствующее сообщение типа "Слово отсутствует в словаре".

В данной версии электронного словаря морфологический анализатор используется в несколько редуцированном виде, например, на выходе слову не приписываются грамматические признаки. Но в будущем для целей анализа словосочетаний "избыточная" сейчас информация будет необходима для синтаксических проверок.

2.3. Система наполнения словаря

При создании самим пользователем индивидуального или корпоративного толкового/терминологического словаря существенной является функция заполнения словаря в части толковых словарных статей для словника выбранной проблемной области. Предполагается, что словник всегда "погружен" в множество представленных в морфологическом словаре слов (их объем эквивалентен нескольким миллионам словоформ).

Организация оболочки комплекса такова, что при заполнении пользователем толковой компоненты словаря не требуется его переинсталляция.

Наполнение словаря может условно включать несколько этапов: первоначальное заполнение базы толковых словарных статей, добавление к уже частично заполненной базе новых статей, замена одной статьи на другую и корректировка текста уже занесенной статьи. Очевидно, что это позволяет динамически пополнять и исправлять базу данных для толковых статей прямо в процессе работы со словарем, поскольку в оболочке словаря незаполненными остаются только memo- поля в таблицах описания толковых статей, а в реально действующем словаре эти поля доступны корректировке самим пользователем.

Для реализации любых этапов наполнения необходим редактор с достаточно развитой стандартной системой команд по записи, исправлению, замене и т.п. Этот редактор также должен предоставлять пользователю средства форматирования текста: задание шрифтового стиля, цветовой подцветки, вставки изображений и т.д.

Проиллюстрируем процесс добавления словарной статьи следующим примером. Необходимо добавить текст толкования слова *изумруд*. Форматирование текста с помощью тэгов языка HTML позволяет задать требуемый стиль отображения:

```
изумр<FONT color=”red”>у</FONT>д <BR>
Прозрачный драгоценный камень ярко-зеленого цвета.
<BR>
<I>Изумрудная вода</I> (цвета изумруда).
```

После запоминания толковой статьи в ответ на запрос о толковании слова *изумруд* этот текст в окошке словаря будет отображен следующим образом (поскольку цвет при печати не отображается, он выделен курсивом):

изумр*уд*
Прозрачный драгоценный камень ярко-зеленого цвета.
Изумрудная вода (цвета изумруда).

3. Реализация комплекса

В связи с принятым решением использовать БД для хранения данных, для реализации системы необходимо выбрать программное средство, позволяющее максимально использовать его возможности. Комплекс как программное приложение должен манипулировать

сравнительно небольшими объемами данных, быть недорогим в эксплуатации и по себестоимости, его целесообразно реализовать в системе позволяющей создавать файл – серверные приложения с расширенными функциями. Оптимальным средством, отвечающим этим требованиям, является Microsoft Visual FoxPro 6.0. Данная средство позволяет в полном объеме реализовать необходимые функции и создать, как конечный продукт, систему, включающую исполняемый модуль, ориентированный для использования в среде Windows 9x/NT.

Microsoft Visual FoxPro 6.0 относится к недорогим средствам разработки приложений для СУБД, позволяет использовать объектную модель программирования и язык запросов SQL, что снижает себестоимость продукта и максимизирует его универсальность и производительность.

Дополнительным средством, призванным упростить создание и универсализировать работу системы является интеграция с WEB Browser'ом. Это решение используется как альтернатива написания программных модулей, анализирующих данные с целью нахождения специальных символов и соответствующего форматирования выводимого пользователю результата (выделения; ударения; задание шрифтов во фрагментах описания толковой статьи). Использование стандартного "просмоторщика" позволяет избежать рутинного программирования, непереносимости данных и вероятности некорректной работы в случае использования специальных символов в рамках текста (как элементов текста). В предлагаемом варианте, вместо специальных символов используются стандартные тэги языка HTML, обеспечивающие широкие возможности в области форматированного вывода текста. Этот язык в будущем позволит добавлять гипертекстовые ссылок, вставлять изображения в словарные статьи и т.п. Он также снимет существующие ограничения по максимальному размеру статьи, которые сегодня составляют 64 Кбайта.

Эксплуатация комплекса возможна на базе любого IBM-совместимого компьютера с оперативной памятью не менее 8 Мб и памятью на жестком диске, определяемой в зависимости от объема морфологической и толковой баз словаря. Для функционирования комплекса требуется операционная система Windows 9x/NT.

4. Инсталляция словаря

Комплекс выполнен в виде самостоятельного пакета, функционирующего в ОС Windows 9x/NT. Стандартный инсталляционный пакет, созданный на базе автоматического инсталлятора "Wise", включает все необходимые файлы и компоненты. Вся настройка системы (регистрация в реестре Windows, добавление программной группы в меню) производится автоматически. Также поставляется программа для полного удаления пакета (удаление ярлыков, исполняемых файлов, БД, динамических библиотек Visual FoxPro, очистка реестра) из компьютера пользователя.

Программа инсталляции не требует от пользователя никаких действий кроме ответов на выводимые на экран вопросы. При запуске файла setup.exe программа предлагает выбрать каталог, в который она будет установлена. После этого начнется собственно процесс инсталляции, т.е. копирование файлов с дискет на жесткий диск. После инсталляции толковый словарь готов к работе.

5. Работа с комплексом

После установки на компьютере комплекса версии 1.0 реализуются следующие режимы работы:

- набор слова в произвольной форме и получение его толкования;
- выделение любого слова из самой толковой словарной статьи и получение его толкования;
- получение для омонимичных словоформ списка лексем-омонимов и выбор из них интересующего пользователя слова;
- получение предыстории работы со словарем;
- добавление толковых словарных статей.

Первые четыре режима относятся собственно к работе со словарем, а пятый – к работе с системой пополнения.

В первом режиме пользователь может набрать слово в любой произвольной форме и получить толковую словарную статью или сообщение об ее отсутствии в словаре. В последнем случае это справедливо либо тогда, когда слово есть, а толкования нет, либо когда слово отсутствует в словаре, либо когда слово введено с ошибками. Экран, иллюстрирующий работу словаря в этом режиме, представлен на рис. 1. (Для демоверсии тексты толковых статей из словаря С.И. Ожегова были подготовлены Ломакиной В.В.).

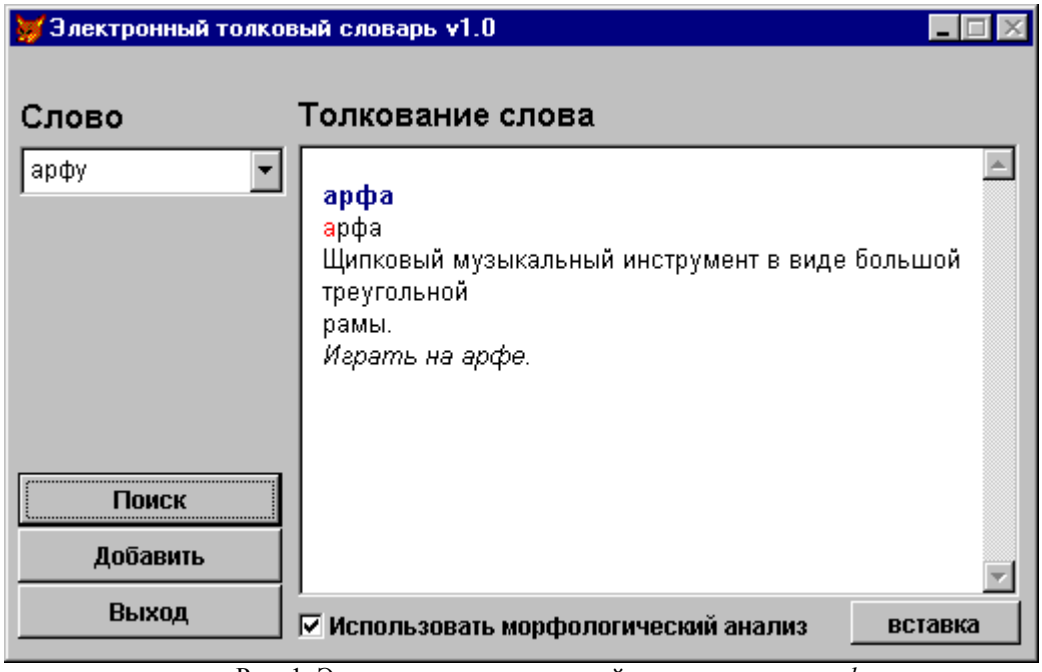

Рис. 1. Экран получения толковой статьи на слово *арфу*.

Во втором режиме пользователь может рассматривать в окошке словаря толковую словарную статью как источник задания интересующих его слов. Для реализации этого режима пользователь выделяет мышкой интересующее его слово (в произвольной его форме) и дает задание на поиск его толкования. В этом случае пользователь работает с толковой словарной статьей как с обычным текстовым файлом.

Третий режим реализуется тогда, когда пользователь задает омонимичное слово, например, *цари*. В результате работы морфологического анализатора на экране появится окошко со списком лексем-омонимов: *царь*, *царить*. В этом списке он может выбрать интересующее его слово и получить его толкование (см. рис.2).

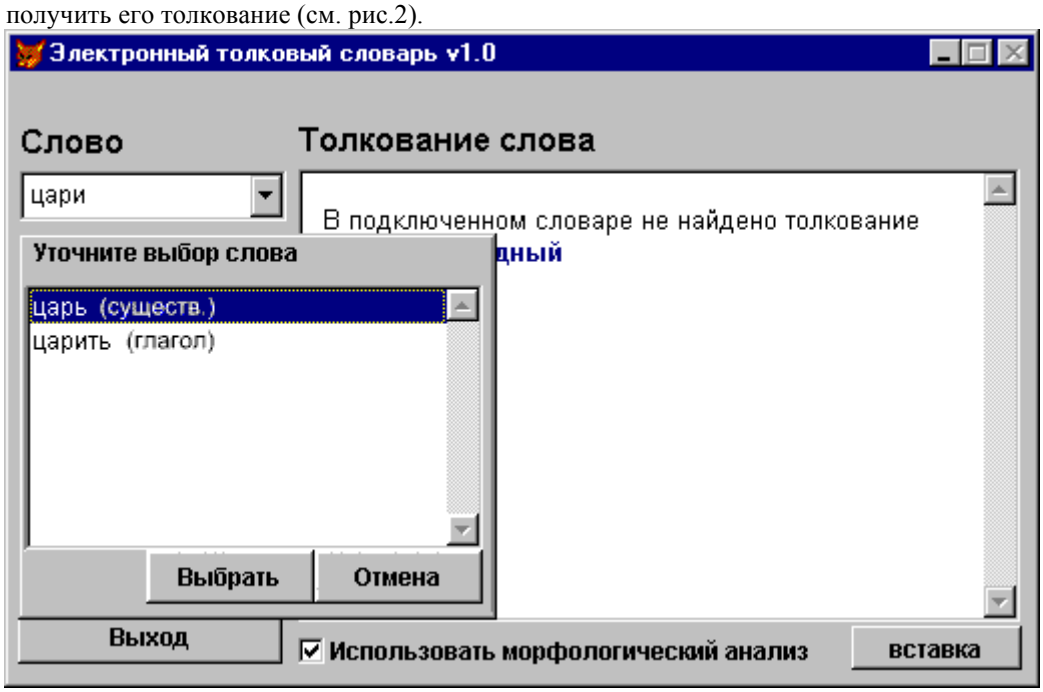

Рис. 2. Экран получения списка лексем-омонимов для словоформы *цари*.

Четвертый режим реализуется в том случае, если пользователь захочет вернуться к любому ранее выбранному им слову. Он может активизировать окошко "предыстории", где хранится список слов ранее анализируемых словарем. В данной версии количество слов ограничивается десятью словами, но этот параметр может быть произвольно увеличен. Такой

режим позволяет пользователю возвращаться на любой из предыдущих этапов работы со словарем, что существенно, например, в процессе обучения. Экран, иллюстрирующий работу словаря в этом режиме, представлен на рис. 3.

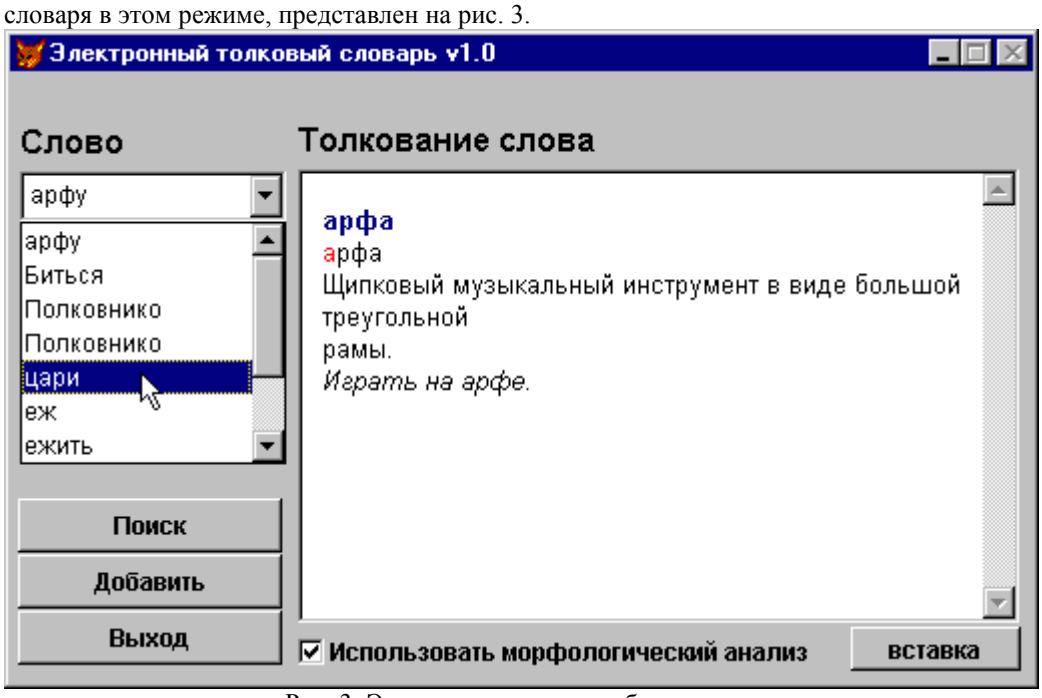

Рис. 3. Экран предыстории работы со словарем.

Пятый режим реализует функцию добавления толкования выбранного слова. Нажатие кнопки «Добавить» в основном окне приложения вызывает окно ввода добавляемого слова и толковой словарной статьи. После ввода текста толковой статьи нажатие кнопки “Добавить” выполняет операции добавления введенной информации в БД. Если к данному слову уже было приписано толкование (т.е. была задана толковая словарная статья), то будет выдано сообщение "Данное слово в словаре уже описано".

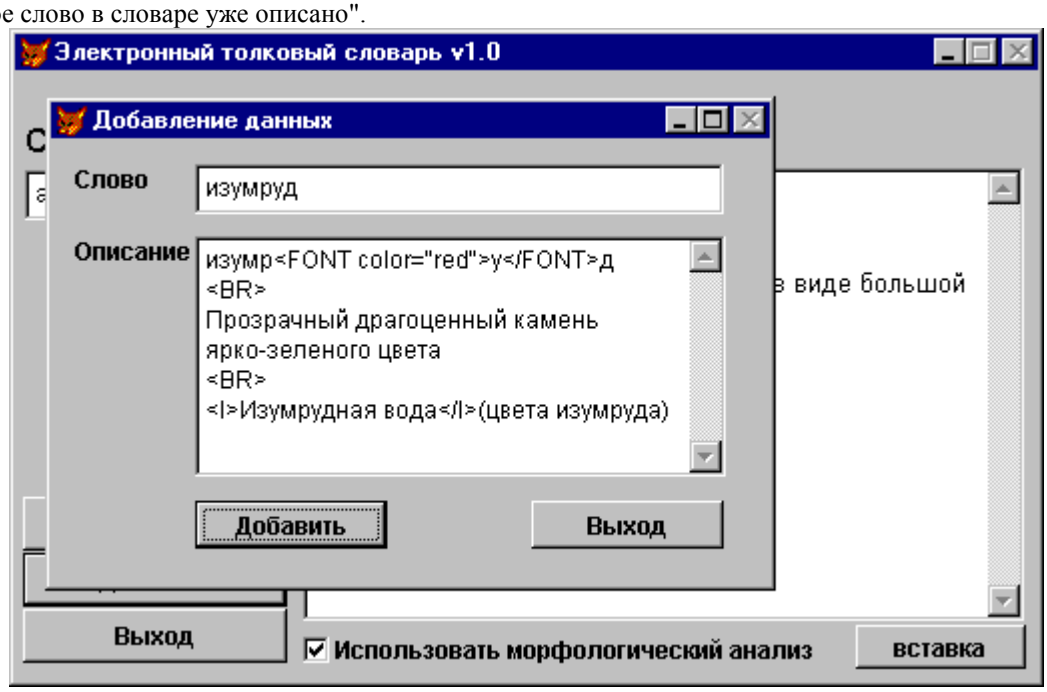

Рис. 4. Экран добавления толковой статьи.

Заключение

Терминологическое обеспечение современных информационных систем отражает потребность в создании различного рода электронных словарей и, в частности, толковых/терминологических словарей для учебного процесса и для практической

профессиональной деятельности. Предлагаемый комплекс лежит в русле этих разработок, тем более, что он обеспечивает возможность адаптации его оболочки к любому словарю толкового типа для различных предметных областей за счет заполнения декларативной компоненты словаря.

Рабочая версия комплекса имеет ряд практических ограничений. Хотя выбранные методы и структура организации комплекса позволяют его расширение без "ломки" принятых решений. В первую очередь к расширениям возможностей комплекса относятся:

- интеграция с текстовым редактором типа Word;
- разработка подробной справочной системы;
- на ввод слова с ошибками или на ввод отсутствующего в словаре слова – выдача списка слов, которые "похожи" на заданное слово;
- ввод в толковые словарные статьи рисунков, графиков и т.п. графической информации, расширяя тем самым толковый словарь до универсального;
- разработка специализированного редактора, позволяющего в вводимой словарной статье использовать простые и стандартные средства "оттенения" выделяемой информации (например, подсветка ударной буквы, выделение шрифтом примеров, форматированием абзацев и т.д.);
- автоматическая идентификация словосочетаний, что потребует привлечение как минимум синтаксического анализатора для распознавания правильно построенных словосочетаний и для получения нормализованных словосочетаний.

Предполагается, что перечисленные выше возможности будут реализованы в следующей версии комплекса.

Литература

Зализняк А.А. Грамматический словарь русского языка. М.: Русский язык, 1980, 879 с.

Лахути Д.Г., Пархоменко В.Ф., Маргаритов Л,И. и др. Русский филолог. Программный пакет. Москва. АО Агама, 1993.

Мальковский М.Г. Диалог с системой искусственного интеллекта. - М.: МГУ, 1985. - 213 с.

Маслов Ю.С. Введение в языкознание. М.: Высшая Школа, 1997, с. 120-124

Поляков А.Е. Электронный словарь языка писателя (на примере языка А.С. Грибоедова). Труды Международного семинара Диалог-99 по компьютерной лигвистике и ее приложениям, в 2-х томах, т. 2. Приложения. Таруса, 1999, с. 230-236.

СЭТС система разработки электронных толковых словарей для русского языка. Каталог выставки "Интеллект-СОФТ'96" - М., НТК "МЕТОД", 1996, с.19.

Хахалин Г.К. Лингвистический транслятор в семействе систем с обработкой ЕЯ-текстов (ретроспекция). Труды VI национальной конференции по Искусственному Интеллекту с международным участием - КИИ-98, 5-11 октября. Пущино, Россия, т.1, 1998, с. 238 - 246.

Complex on development of the individual and/or corporative electronic explanatory dictionaries

Gennady Khakhalin, Nikolay Bogdanov, Sergey Platonov
khakhalin@got.mmtel.ru

Complex on making the explanatory dictionaries ensures a development and support of user electronic dictionary with filling on any application domain. Complex consists of shells of electronic dictionary with the access to the dictionary entry on any word form and filling systems on the concrete application domain. It serves an user or group of users by the facility for making individual and/or corporative dictionaries. The Electronic explanatory dictionaries realized by means of the given complex can be used in the process of decision making or in the process of education. Dictionary ensures the adaptation possibility of the electronic shell to any dictionary filling the declarative components of dictionary. Variety of modes allows: enter any word in the free form and get its

interpretation, not conceiving on the standard form; get the whole history of the dictionary functioning, returning on any one of previous stages; to work with the dictionary article as with the usual text file and etc. Known electronic explanatory dictionaries "wired" and do not give a chance user to create "own" dictionary. Other electronic dictionaries are bilingual and designed on several other technologies.

Key words: linguistical dictionaries, electronic explanatory dictionary, linguistical database, morphological analysis, word form normalization, tuning on the application domain, educating facility, individual and/or corporative electronic dictionary