УДК 519.7

**А.С. Пузик**
ХНУРЭ, Украина, as308@mail.ru

# ЛЕКСИКОГРАФИЧЕСКАЯ СИСТЕМА ЭЛЕКТРОННОГО ТЕРМИНОЛОГИЧЕСКОГО СЛОВАРЯ

Статья посвящена описанию программной системы русско-украинско-английского терминологического словаря. В статье описаны подходы к первичной обработке исходных данных. Описана организация модели данных для трехъязычного словаря и перспективы развития модели при переходе к многоязычным словарям. Приведена база данных и архитектура программной системы. Показаны базовые принципы работы со словарем с точки зрения пользователя.

ЭЛЕКТРОННЫЙ СЛОВАРЬ, АЛГЕБРА КОНЕЧНЫХ ПРЕДИКАТОВ, OFFICE AUTOMATION, БАЗА ДАННЫХ, ПАРСИНГ, ШАБЛОН MVVM

## Введение

Компьютерные системы давно и прочно вошли в наш мир. На них основаны как простые вебсайты, так и сложные информационные порталы. С конца прошлого века начала свое развитие компьютерная лексикографии. Ручной труд лексикографа переводится в электронную форму, на смену бумажным словарям приходят электронные, которые доступны в том числе и через интернет. Электронные словари могут использоваться как база для дальнейшего развития систем интеллектуальной обработки естественной речи, так и в качестве средства получения актуальной переводной информации. Кроме того электронные словари, в отличие от бумажных, позволяют вносить исправления по мере необходимости, без больших затрат, тем самым позволяя оперативно реагировать на возможные недоработки, изменения в языке, тем самым поддерживая актуальность словаря.

В данной статье описывается подход к построению программной системы электронного трехъязычного терминологического словаря по информатике и радиоэлектронике. Описаны проблемы, связанные с представлением, обработкой и хранением данных, приведена лексикографическая база данных и описана внутренняя архитектура системы. Кроме того приведены примеры работы с программной системой.

## 1. Постановка проблемы и цели разработки

Построение электронного словаря является трудоёмкой процедурой состоящей их нескольких шагов. При построении программной системы трехъязычного словаря необходимо учитывать особенности описания типов и структур данных в базе данных, способы изменения состояний данных и способы извлечения данных, а также аспект их целостности. Часть подходов к решению этих проблем решаются при помощи средств лексикографических систем [1, 2] и алгебры конечных предикатов [3, 4].

При построении программной системы электронного словаря в качестве исходных данных использовались предварительно отсканированные и распознанные страницы из двуязычного русско-украинского словаря по информатике и радиоэлектронике. Одной из проблем преобразования, является корректура отсканированных текстов. В данной работе требуется показать подходы к обработке отсканированных текстов словарных статей, построить адекватную модель данных для трёхъязычного словаря и реализовать программную систему для трехъязычного терминологического русско-украинско-английского электронного словаря по информатике и радиоэлектронике.

## 2. Обработка отсканированных текстов

Создание электронных словарей состоит их нескольких этапов. Сначала сканируются и распознаются словарные статьи, производится коррекция артефактов распознавания. Следующим шагом полученный текст разбивается на массив отдельных словарных статей, а потом производится их декомпозиция по формальным признакам [2].

Документы формата «doc» были исходными для данной работы. Формат «doc» - это бинарный формат файлов, который используется в программе MSWord.

Исходный двуязычный словарь построен на основе алфавитно-гнездового принципа [5]. Русское слово-термин является заголовочным словом. Гнездо включает терминологические словосочетания, элементом которых является заголовочный термин. Заголовочное слово заменяется в терминологических словосочетаниях на тильду, а сами терминологические словосочетания строятся таким образом, чтобы тильда была на первом месте. На основании этой информации можно извлечь нужные данные из отсканированных файлов.

Формат «doc» имеет ряд особенностей, из-за которых напрямую использовать данные документов

довольно затруднительно. Разные секции документов содержат распознанный текст, который надо сгруппировать по определенным признакам. Также в тексте содержится довольно большое количество артефактов распознавания, которые надо исправить для дальнейшей обработки информации (рис. 1).

Одним из способов получения необходимой информации является доступ к содержимому документов при помощи технологии Office COM Automation посредством VBA[6]. Этот способ является наиболее простым подходом, поскольку MSWord предоставляет программные интерфейсы для парсинга документов.

При подходе к созданию многоязычных словарей необходимо учитывать такой аспект представления данных, как кодировка. Существует формат кодирования юникод (Unicode), который позволяет представлять символы любого языка в едином формате. Таким образом после обработки документов MSWord, данные были извлечены и сохранены в текстовый юникодный формат. В получившихся файлах хранилась только информация о терминах без учета форматирования, что облегчило их дальнейшую обработку.

Неправильно распознанные символы, знаки переносов, пустые строки, латинские буквы вместо кириллицы, неправильные скобки являлись основной проблемой корректуры. Однако в этих ошибках были определенные закономерности, что позволило организовать их исправление в автоматизированном режиме при помощи регулярных выражений.

Пример из разбитого на переводы строк, но необработанного файла:

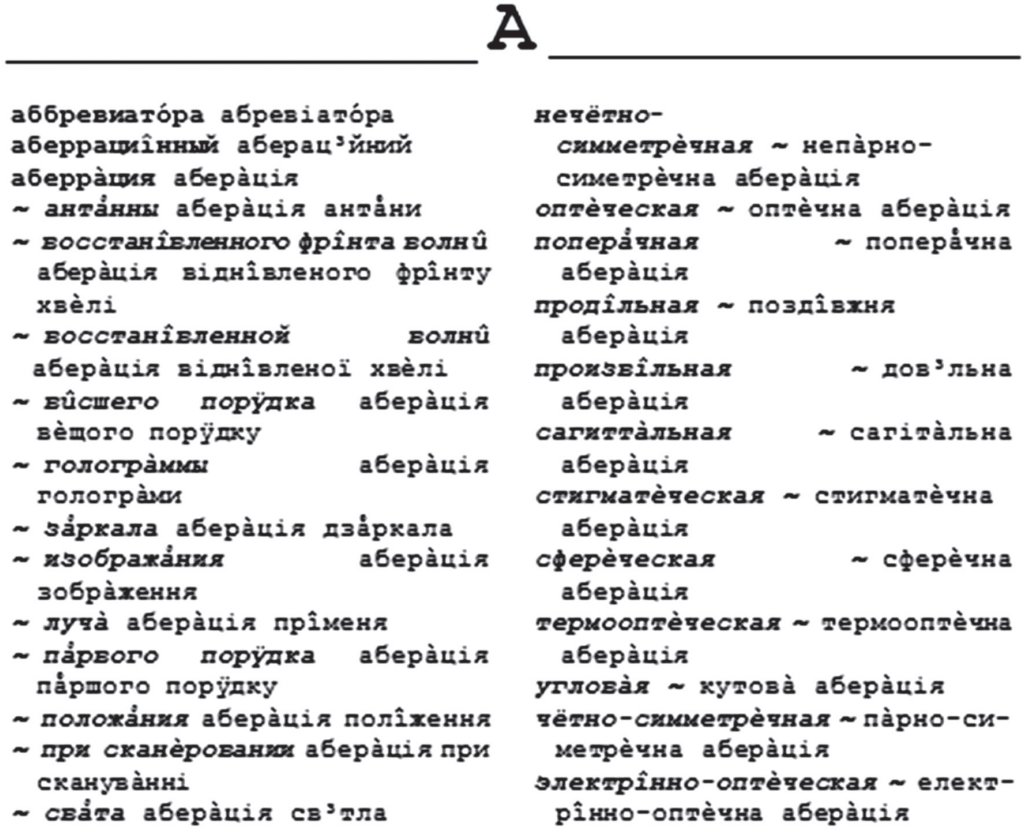
...
*1.(моно, не)хроматическая (*
*2.моно, не)- хроматична аберація)*
...
*1.( -виток*
*2.ампер-виток*
...

В первом случае скобка из верхней строки должна принадлежать нижней, во втором скобка – артефакт распознавания.

Для заполнения внутренних структур словаря при дальнейшей обработке были выбраны такие характеристики терминов, как отрасль знаний, семантика, изменяемая часть слов и прочее.

### 3. Организация модели данных трехъязычного словаря

Обработка, хранение и представление пользователю являются основными проблемами при построении программной системы. Сложная структура лингвистического материала является одной из причин, которые возникают при создании электронного словаря. Частично проблемы и подходы



**А**

**аббревиато́ра** абревіато́ра
**аберрацио́нный** аберац'йний
**аберра̀ция** аберà̀ція
**~ анта́нны** аберà̀ція анта́ни
**~ восстано́вленного фро́нта волно́й** аберà̀ція віднóвленого фрóнту хвèлі
**~ восстано́вленной волно́й** аберà̀ція віднóвленої хвèлі
**~ вы́сшего поря́дка** аберà̀ція вèщого поря́дку
**~ гологра̀ммы** аберà̀ція гологра̀ми
**~ за́ркала** аберà̀ція дза́ркала
**~ изображе́ния** аберà̀ція зобрàження
**~ лучà** аберà̀ція прóменя
**~ пе́рвого поря́дка** аберà̀ція пе́ршого поря́дку
**~ положе́ния** аберà̀ція полóження
**~ при сканѝровании** аберà̀ція при скануванні
**~ сва́та** аберà̀ція св'тла
**нечётно-симметрѝчная ~** непàрно-симетрèчна аберà̀ція
**оптѝческая ~** оптѝчна аберà̀ція
**попере́чная ~** попере́чна аберà̀ція
**продóльная ~** поздóвжня аберà̀ція
**произвóльная ~** дов'льна аберà̀ція
**сагитта̀льная ~** сагіта̀льна аберà̀ція
**стигматѝческая ~** стигматѝчна аберà̀ція
**сферѝческая ~** сферѝчна аберà̀ція
**термооптѝческая ~** термооптѝчна аберà̀ція
**угловà́я ~** кутовà аберà̀ція
**чётно-симметрѝчная ~** пàрно-симетрèчна аберà̀ція
**электрóнно-оптѝческая ~** електрóнно-оптѝчна аберà̀ція

**Рис. 1. Пример входных данных**

к их решению описаны для двуязычных словарей и лексикографических систем в целом [1, 2].

Двуязычный словарь можно представить в виде набора переводных эквивалентов для каждого термина. При этом некоторые термины могут иметь многозначную семантику, соответственно при построении связи это надо учитывать. Когда один из языков является основным, то переводные эквиваленты приводятся относительно этого языка. В случае двустороннего перевода появляется второй аналогичный список для другого языка. При этом семантика самого термина становится размытой между переводными эквивалентами языков.

В базе данных электронного словаря переводные эквиваленты и связи между ними будут храниться в соответствующих таблицах. При этом для омонимичных терминов количество связей будет увеличиваться, а выделение семантики термина будет являться дополнительной задачей, которая потребует дополнительных усилий для решения. При переходе к многоязычным словарям, особенно, когда планируется, свободное переключение между языками, появляется проблема увеличения количества связей пропорционально количеству языков. Это случай связи многие ко многим.

Для решения данной проблемы предлагается введение дополнительного уровня косвенности. Им будет являться абстракция, обозначающая семантику термина. Это позволит перейти от отношения многие ко многим к отношению один ко многим (рис. 2). Также при таком подходе появляется привязка термина к семантике, что открывает возможность использовать контекст для перевода текстов. При дальнейшем развитии словаря терминам можно будет добавлять толкования и семантически однозначные примеры использования.

Таким образом база данных электронного словаря будет содержать таблицу терминов и таблицу переводных эквивалентов. Благодаря такому подходу останется возможность получения всех переводов термина, включая семантически разные, путем простой выборки из таблицы переводных эквивалентов.

Данный подход к построению электронного словаря позволяет перейти от двуязычного словаря к многоязычному словарю.

### 4. Описание лексикографической базы данных и архитектуры программной системы

На основе вышеизложенного материала была построена лексикографическая база данных трехъязычного словаря (рис. 3).

Таблица Conception:

а) Id – идентификатор термина;

б) ParentId – идентификатор родительского термина, используется для терминологических словосочетаний;

в) Semantic – идентификатор семантики термина;

г) Topic – идентификатор отрасли знаний термина.

Таблица Description:

а) Id –идентификатор переводного эквивалента;

б) Description – написание переводного термина в именительном падеже, единственного числа на любом из языков;

в) ConceptionId – идентификатор термина, к которому относится переводной эквивалент;

г) Language – идентификатор языка, к которому относится переводной эквивалент;

д) PartOfSpeech – идентификатор части речи, для терминов, отличных от существительных;

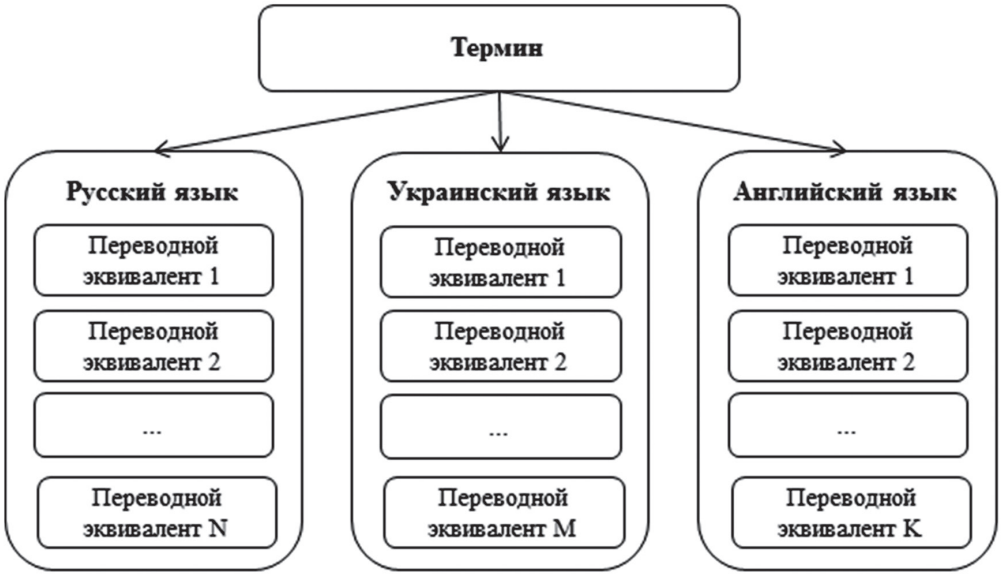



**Рис. 2. Предлагаемая модель хранения данных**

е) ChangableType – идентификатор изменяемой части слова (например род. - родительный падеж, мн. – множественное число);

ж) ChangablePart – написание изменяемой части речи (например окончание в родительном падеже «-та»);

Таблица Semantic:

а) Id – содержит только идентификатор семантики, в дальнейшем будет расширена.

Таблица SemanticTranslation:

а) Id – идентификатор перевода семантики;

б) SemId – идентификатор семантики, для которой этот перевод;

в) LangForId – идентификатор языка, для которого этот перевод;

д) Translation – непосредственно сам перевод на языке LangForId.

Таблица Topic:

а) Id – содержит только идентификатор отрасли знаний, в дальнейшем будет расширена.

6. TopicTranslation

а) Id – идентификатор перевода отрасли знаний;

б) TopicId – идентификатор отрасли знаний, для которой этот перевод;

в) LangForId – идентификатор языка, для которого этот перевод;

д) Translation – непосредственно сам перевод на языке LangForId.

Таблица PartOfSpeech:

а) Id – содержит только идентификатор части речи, в дальнейшем будет расширена.

Таблица PartOfSpeechTranslation:

а) Id – идентификатор перевода части речи;

б) PartOfSpeechId – идентификатор части речи, для которой этот перевод;

в) LangForId – идентификатор языка, для которого этот перевод;

д) Translation – непосредственно сам перевод на языке LangForId.

Таблица ChangablePartType:

а) Id – содержит только идентификатор изменяемой части слова, в дальнейшем будет расширена.

Таблица ChangablePartTypeTranslation:

а) Id – идентификатор перевода изменяемой части слова;

б) PartOfSpeechId – идентификатор изменяемой части слова, для которой этот перевод;

в) LangForId – идентификатор языка, для которого этот перевод;

д) Translation – непосредственно сам перевод на языке LangForId.

Таблица Languages:

а) Id –идентификатор языка, используемого в словаре;

б) Name – название языка по умолчанию.

Таблица LangTranslation:

а) Id – идентификатор языка, используемого в словаре;

б) LangIdNeeded – идентификатор языка, который переводится;

в) LangForId – идентификатор языка, для которого этот перевод;

д) Translation – непосредственно сам перевод на языке LangForId.

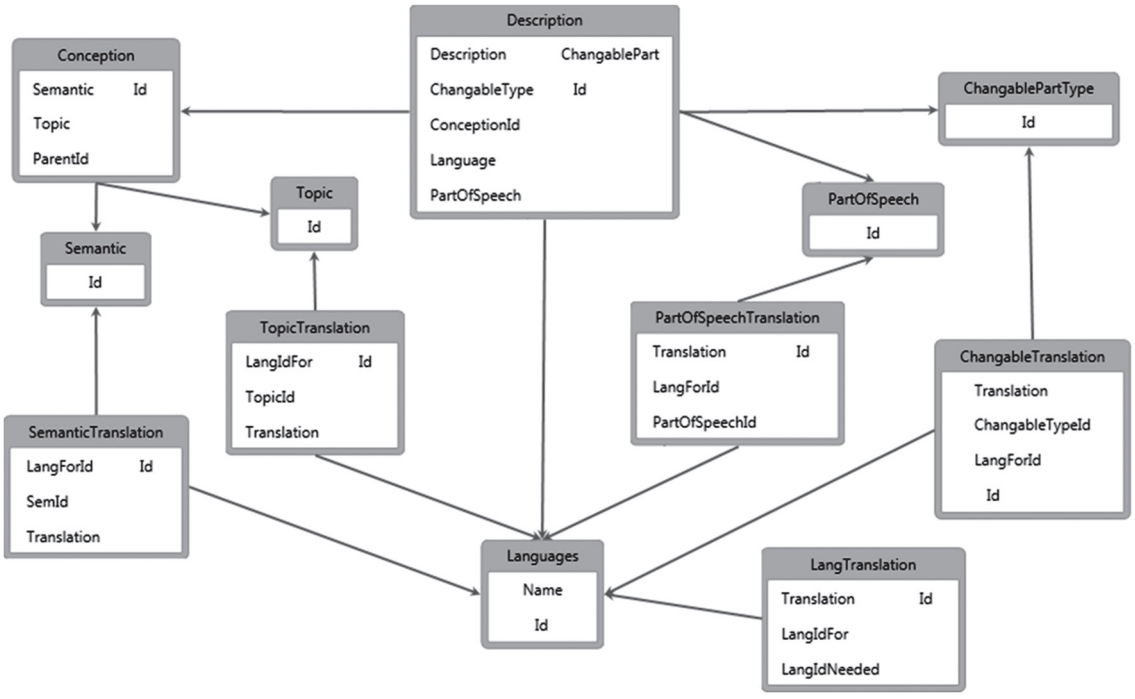



**Рис. 3. Лексикографическая база данных трехъязычного словаря**

Для построения программной системы электронного словаря был использован шаблон проектирования MVVM (Model-View-ViewModel, Модель–Представление–Модель представления) (рис. 4).

Модель – это термин обозначающий классы реализации внутренних данных и логику взаимодействия между ними в программной системе. Представление – это термин, который обозначает классы, обеспечивающие взаимодействие пользователя с графическим интерфейсом. Модель представления – это посредник между моделью и представлением. Модель представления с одной стороны обрабатывает логику представления данных, а с другой стороны передает информацию о действиях пользователя в модель.

Благодаря такому подходу интерфейс пользователя отделён от основной логики программы, что позволяет вносить изменения в различные компоненты программной системы независимо друг от друга.

### 5. Примеры использования программной системы трёхъязычного словаря

Программная система трёхъязычного словаря предназначена для создания, редактирования, просмотра терминов и их переводных эквивалентов на русском, украинском и английском языках (рис. 5).

Основной язык выбирается в выпадающем списке сверху по центру главного окна. В левой панели находятся все переводные термины сгруппированные по алфавиту основного языка. Словарь поддерживает функцию поиска, для этого в окне «Поиск» надо ввести интересующую часть слова. Поиск может осуществляться как с учетом ударения, когда ударная буква помечается символом «#», так и без него. После нажатия кнопки «Найти» в левой панели выбирается термин отвечающий первому найденному совпадению.

Центральная панель отображает переводные эквиваленты для всех языков для выбранного термина.

Правая панель предназначена для редактирования терминов. Для добавления переводного эквивалента в поле «Введите текст» надо ввести текст переводного эквивалента термина для выбранного языка. После этого надо нажать кнопку «Добавить» в секции «Переводной эквивалент». Для изменения существующего переводного эквивалента, его надо выбрать в центральной панели, в поле «Введите текст» обновить значение и нажать кнопку «Изменить». Для удаления переводного эквивалента надо его выбрать в центральной панели и нажать кнопку «Удалить».

Кроме того для переводных эквивалентов может быть указана часть речи в выпадающем меню «Часть речи». Если необходимо показать изменение фонем или ударения, то в поле «Изменяемая часть» вводится текст изменяемой части. Это может быть либо слово целиком (например, для слова «вісь» в этом поле будет «осі»), либо другая изменяемая часть (например, для слова «дейтрон» надо ввести «-на», а для слова «нелінійність» – «-ності»). В поле «Тип изменяемой части» вводится тип изменяемой части (например, родительный падеж «род.»)

Для термина в целом можно ввести дополнительный семантические пометы. Условное обозначение отрасли в которой применяется термин (например, мат., физ. и т.д.) выбирается в выпадающем меню «Тема». В меню «Семантика» выбирается краткое толкование семантики термина (например, слово «акт» может иметь два значения «действие» и «документ»). После нажатия кнопки применить к термину данные из этих полей вносятся в базу данных. При нажатии кнопки «Добавить термин» пользователю будет представлено окно,

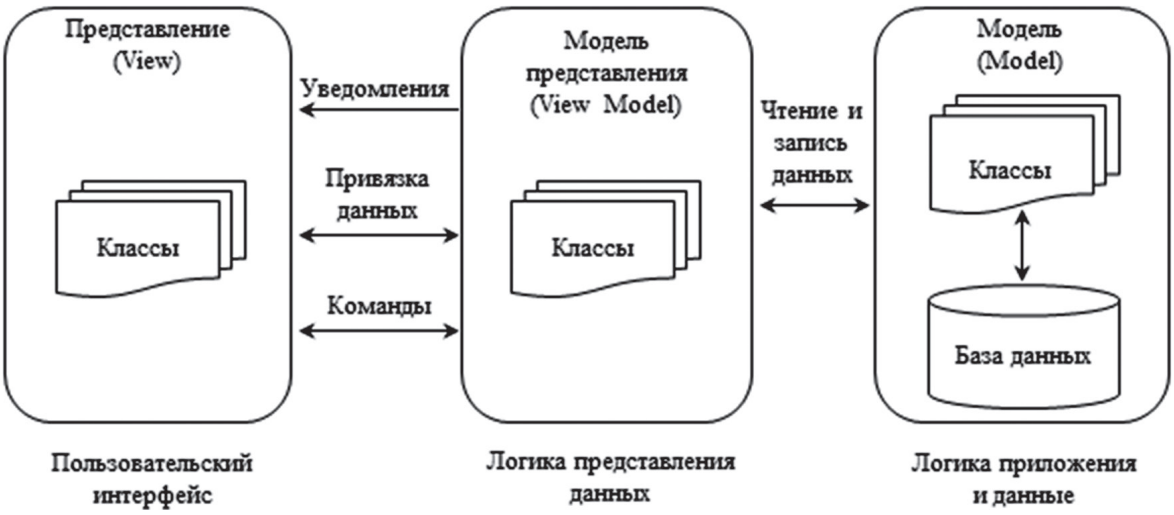



**Рис. 4. Архитектура программной системы**

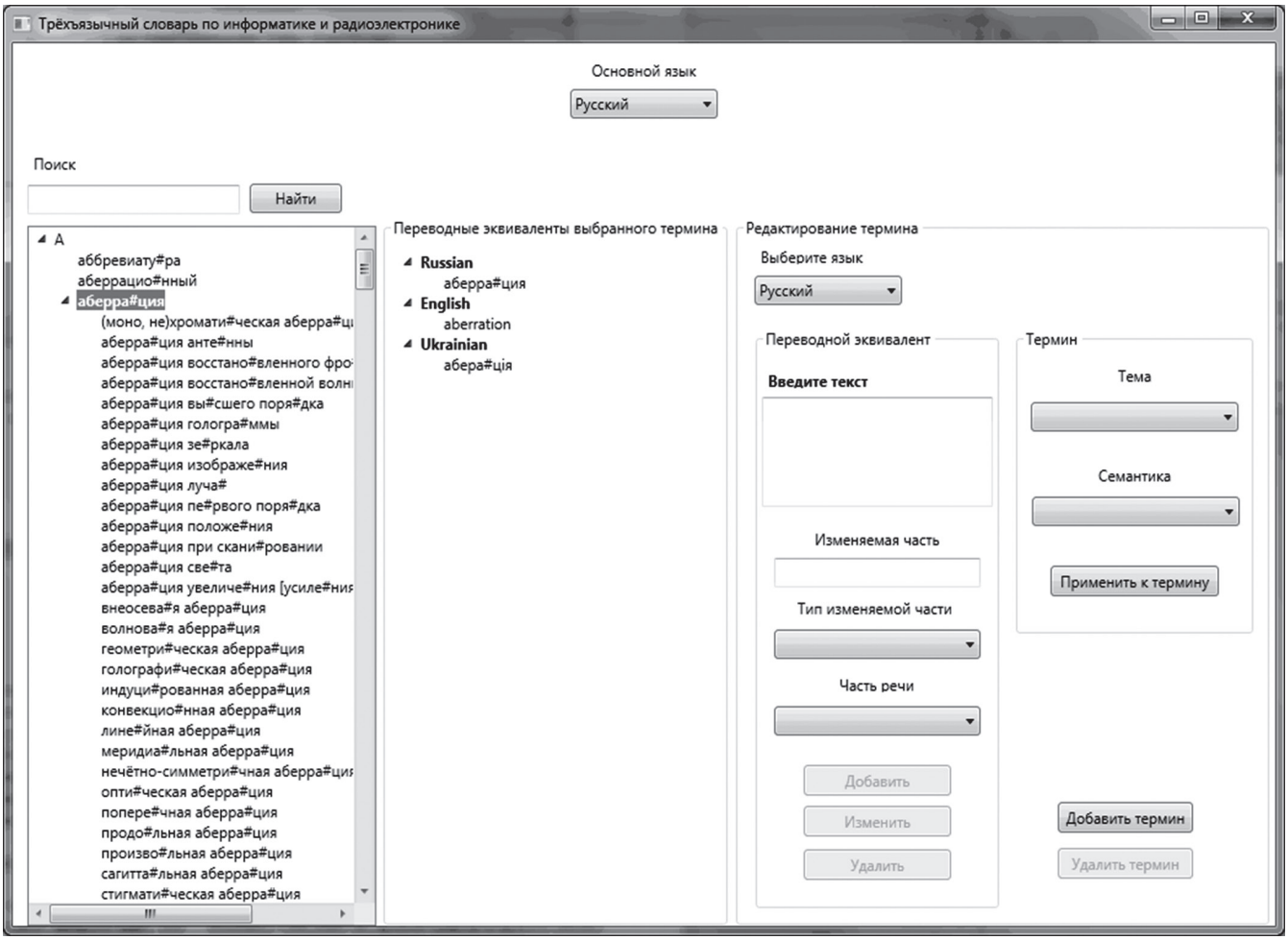



**Рис. 5. Главное окно**

в котором можно добавить термин, заполнив необходимые поля. При нажатии кнопки «Удалить термин», выбранный в левой панели термин будет удален, со всеми переводными эквивалентами.

### Выводы

Таким образом, была создан электронный трехъязычный терминологический русско-украинско-английский словарь по информатике и радиоэлектронике. При его создании были использованы подходы, позволяющие избежать ранее описанных проблем, возникающих при построении электронных словарей [1, 5]. В частности при помощи регулярных выражений удалось откорректировать входные данные. При помощи математического аппарата теории лексикографических систем и алгебры конечных предикатов [7] удалось избежать избыточности в решениях. Предложенный подход к построению модели данных словаря можно также применить и для многоязычных словарей.

В статье приведено детальное описание лексикографической базы данных словаря и архитектуры программной системы. Приведены примеры использования словаря пользователем.

Одним из достоинств системы является изначальное равноправие языков, что позволяет назначать основной язык в зависимости от текущих потребностей.

### Список литературы

**1.** *Широков В.А.* Комп'ютерна лексикографія. – Київ: науково виробниче підприємство «Видавництво «Наукова думка» НАН України», 2011. – 352 с. **2.** *Рабулець О.Г., Широков В.А., Якименко К.М.* Дієслово в лексикографічній системі – К.: Довіра, 2004. – 259 с. **3.** *Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.* Теория интеллекта: учеб. – Харьков: Изд-во СМИТ, 2006. – 571 с. **4.** *Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.* Мозгоподобные структуры: справочное пособие. Том первый – К.: Наукова думка, 2011. – 460 с. **5.** *Остапова И.В.* Лексикографическая структура этимологических словарей и их представление в цифровой среде // Прикладная лингвистика и лингвистические технологи: сборник научных трудов. – 2007. – С. 236-245. **6.** Word VBA reference // [Електр. Ресурс]. – Режим доступу: https://msdn.microsoft.com/en-us/library/office/ee861527.aspx **7.** *Вечирская И.Д.* Разработка трехязычного терминологического словаря на основе алгебры конечных предикатов// Бионика интеллекта: науч.-техн. журнал. – 2011. – № 2(76). – С. 109-113.

Поступила в редколлегию 17.11.2016.