Разработка системы для просмотра и редактирования тезаурусов на основе XML/Java2/Oracle9i технологий

А. Сухоногов (Петербургский государственный университет путей сообщения),
С. Яблонский (Петербургский государственный университет путей сообщения, ЗАО “Руссикон”)

Введение

Сегодня тезаурусы и построенные по их типу словари (например, лексиграфическая база данных WordNet Принстонского университета [1]) широко используются для

- поиска в Интернете,
- автоиндексации документов,
- как одна из эффективных форм представления знаний для Semantic Web и пр.

Европейские проекты EuroWordNet (закончен) и BalkanNet (продолжается) сделали возможным работу с WordNet практически на всех основных европейских языках. При этом связь различных языковых версий WordNet осуществляется через межязыковой индекс (Inter-Lingual-Index − ILI), общий для всех версий [5]. Таким образом, сегодня уже существует многоязычный WordNet общего назначения. Все это определяет необходимость создания русской версии WordNet и подключения ее через ILI к уже существующим европейским версиям. Работы в этой области ведутся в МГУ, СПГУ, ЗАО “Руссикон” и в ряде других центров.

Практика работы с тезаурусами и с WordNet, в частности, показала, что хорошие результаты при поиске по запросам (анализе текстов) общей тематики не исключают в ряде случаев неудовлетворительные результаты при работе с техническими и другими специальными текстами. Это обусловило задачу разработки дополнительных помет для определения предметных или проблемных областей в WordNet (domain labels), отражающих не общую, а предметно или ситуативно определенную специфику той или иной области [2]. При построении такого дополнения (расширения) Wordnet, можно более эффективно использовать методы тезаурусной обработки текста запроса и текстов ПО для получения более точного результата.

В настоящей работе рассматривается разработка программной системы для создания, просмотра и редактирования многоязычных версий WordNet и, в более широком смысле, − тезаурусов для произвольных проблемно-предметных областей. Эта программа используется нами для создания параллельной англо-русской версии WordNet, ряда проблемно-предметных расширений WordNet, а также предметных классификаторов.

Для создания русской версии WordNet применяются лингвистические ресурсы и программное обеспечение ЗАО “Руссикон” [6; www.russicon.ru].
В настоящее время разработаны две версии системы:

- для работы в локальной сети (Delphi + СУБД Oracle 9i);
- Интернет/Интранет версия (Java Server Pages + СУБД Oracle 9i).

В базе данных хранится расширенный англо-русский вариант WordNet. Разработанный редактор позволяет вести неограниченное количество языковых версий WordNet (тезаурусов), осуществлять навигацию, поиск и редактирование каждой языковой версии и любой языковой пары.

Организация WordNet

Базовой словарной единицей в WordNet является не отдельное слово, а так называемый синсет. Аналогом этого термина в отечественных источниках можно считать такие понятия, как класс эквивалентности [9,10] или семантический класс. Наглядно, понятие синсета может быть представлено лексической матрицей, рис.1.

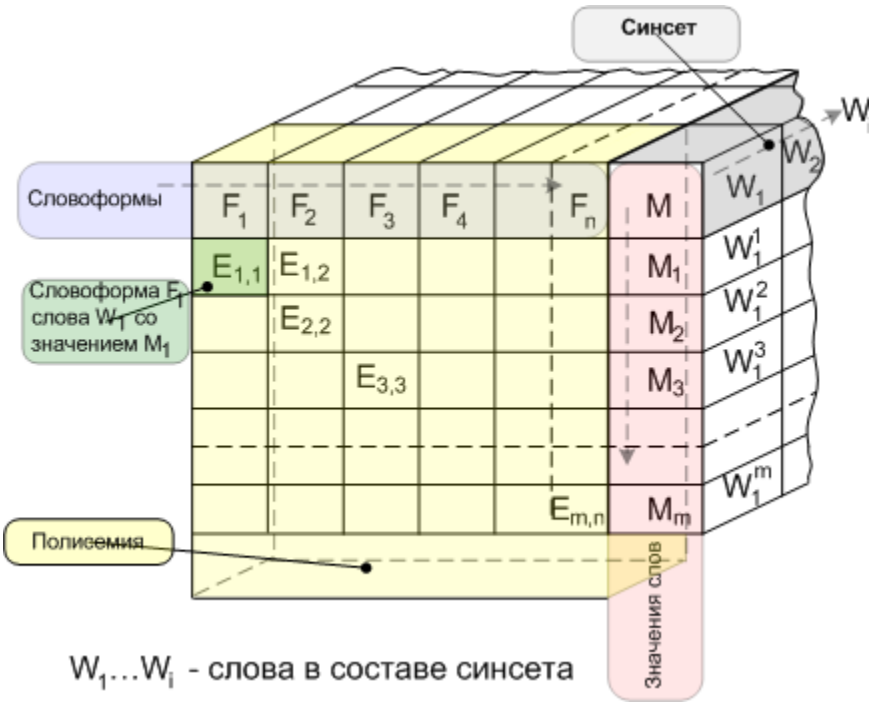
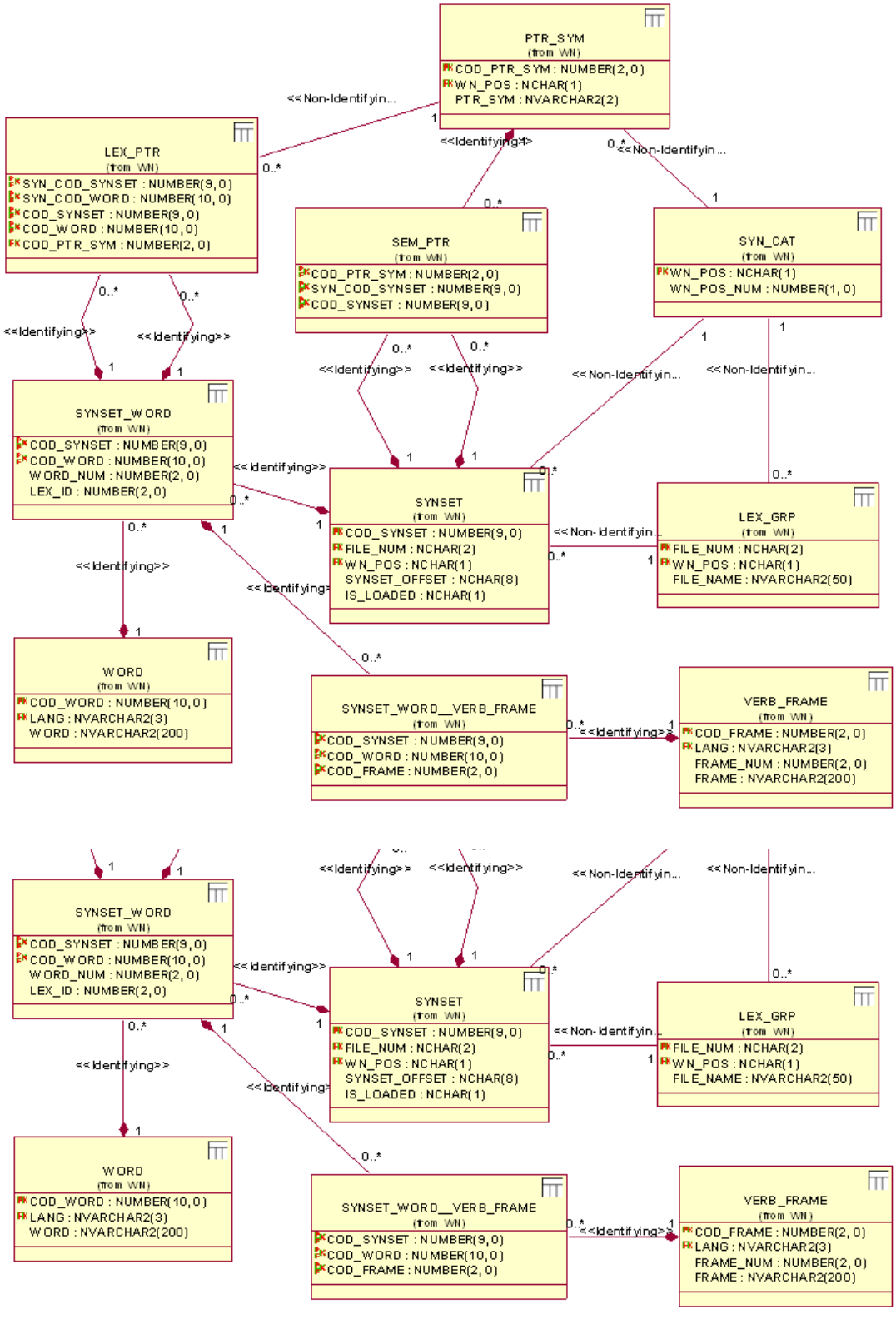


Рис.1 Синсет

В виде синсетов (synset) представлены группы существительных, глаголов, прилагательных и наречий, между которыми определяются различные отношения. Именованные связи между словами синсетов позволяют выделять антонимы для заданного слова. Каждый синсет содержит список синонимов или синонимичных словосочетаний и указатели, описывающие отношения между ним и другими синсетами. Эти указатели позволяют устанавливать различные отношения: иерархические, соответствующие иерархиям род/вид, часть/целое; не иерархические, в частности определяющие атрибуты для значений некоторых синсетов, например, вес = {большой, средний, маленький}. Слова в синсетах логически сгруппированы так, что могут быть взаимозаменяемыми в некотором контексте. Для каждого синсета однонозначно определяется его положение в грамматической и лексической классификации. Слово или словосочетание может появляться более чем в одном синсете и иметь более одной категории части речи.

Слова, для которых существует омонимия и полисемия одновременно включаются в несколько синсетов и могут быть причислены к различным синтаксическим и лексическим классам. Каждый синсет сопровождается словесным описанием его значения, не допускающим его неоднозначного прочтения и понимания.

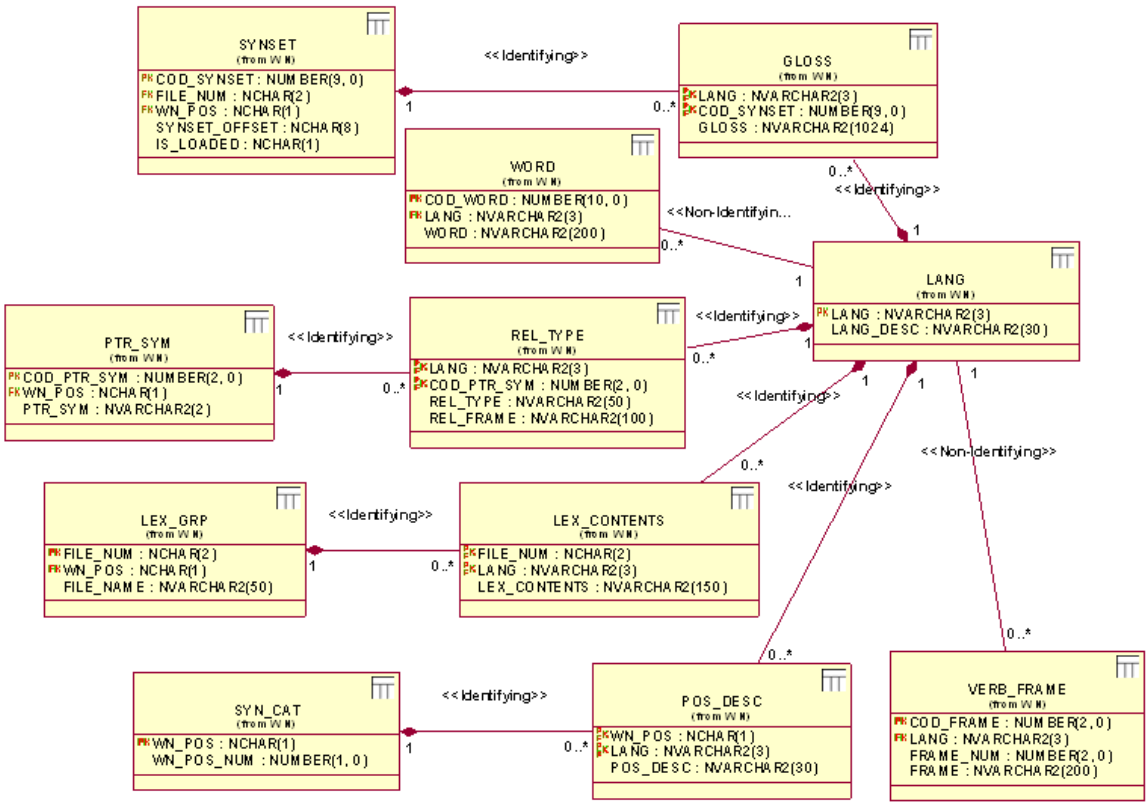
Рис. 2.а UML-диаграмма классов WordNet



UML, RDF и XML модели WordNet и классификаторов (рубрикаторов и тезаурусов)

Для построения концептуальной и логической модели данных базы данных WordNet и других подобных объектов построены Unified Modeling Language (UML) [4], Resource Definition Framework (RDF) [3] и XML модели.

Так, в среде Rational Rose разработаны UML-диаграммы классов WordNet. Эта диаграмма имеет два представления. На рис. 2а представлены основные отношения между классами в WordNet, предложенные в рамках проекта Princeton WordNet (http://www.cogsci.princeton.edu/~wn). На диаграмме рис.2б представлены все зависимые от языка представления части классов основной диаграммы. Также разработана диаграмма классов ряда классификаторов (рубрикаторы и тезаурусы) с различными отношениями в соответствии с ГОСТ 7.49-84 [9] и ГОСТ 7.25-2001 [10] (рис. 3). Дополнительно к требованиям стандарта реализовано представление данных классификаторов на различных языках, а также возможность реализации перекрестных отношений между классификаторами.



На основе данных UML-диаграмм сгенерированы соответствующие логические схемы модели данных для СУБД Oracle9i.

На их основе разработаны два варианта реализации базы данных (БД) Oracle 9i:

- в виде традиционной реляционной БД;
- в виде XML БД [7].

Рис. 2.б UML-диаграмма классов WordNet. Языковые ресурсы

Таблица 1. Наименования классов в модели WordNet

Наименование	Код
Язык представления	LANG
Лексические группы WordNet	LEX_GRP
Лексический указатель	LEX_PTR
Название части речи	POS_DESC
Определения и примеры	GLOSS
Семантический указатель	SEM_PTR
Слово	WORD
Синсет	SYNSET
Синтаксическая категория	SYN_CAT
Содержание лексической группы	LEX_CONTENTS
Фрейм глагола	VERB_FRAME
Указатели	PTR_SYM
Тип связи по указателю	REL_TYPE
Элемент синсета	SYNSET_WORD

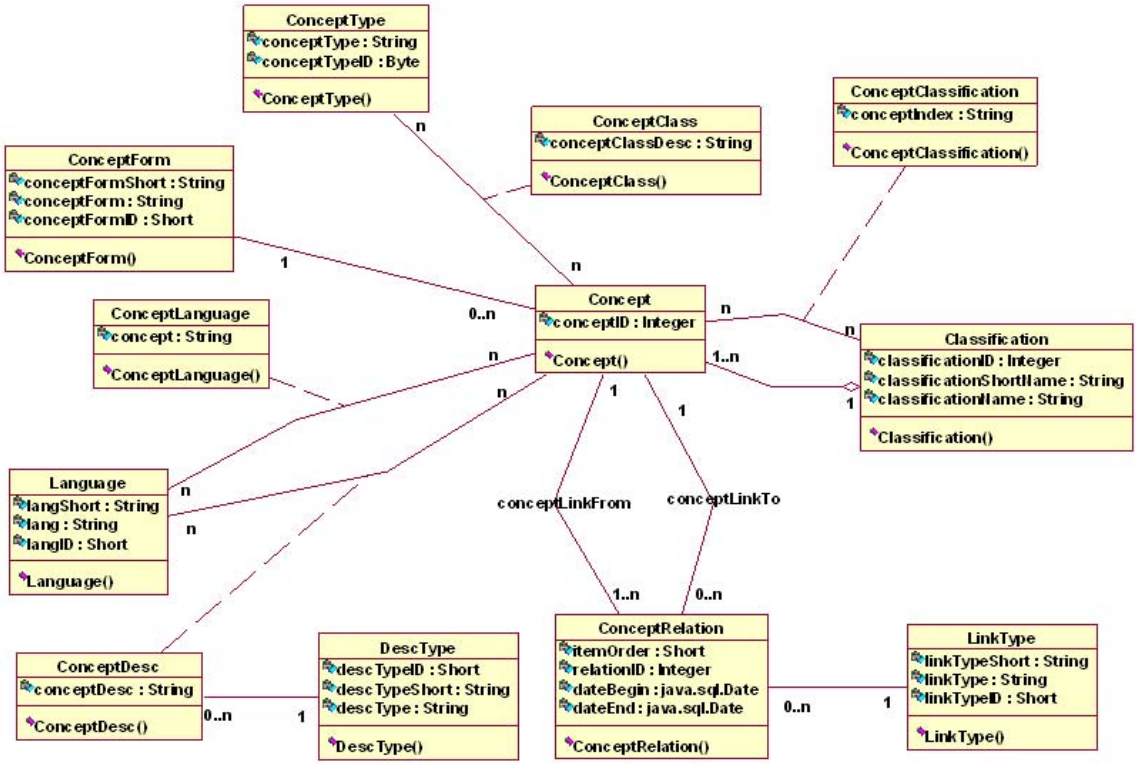
Таблица 2. Наименования атрибутов классов в модели WordNet

Атрибут	Код
Язык представления	LANG
Признак формирования синсета	IS_LOADED
Идентификатор в лексическом файле	LEX_ID
Код слова	COD_WORD
Код синсета	COD_SYNSET
Код синтаксической категории	WN_POS
Код фрейма	COD_FRAME
Код указателя	COD_PTR_SYM
Имя файла	FILE_NAME
Определение и примеры	GLOSS
Описание в WordNet	POS_DESC
Наименование языка представления	LANG_DESC
Номер категории	WN_POS_NUM
Номер слова в синсете	WORD_NUM
Номер фрейма	FRAME_NUM
Номер файла	FILE_NUM
Офсет синсета	SYNSET_OFFSET
Слово	WORD
Символ указателя в WordNet	PTR_SYM
Содержание	LEX_CONTENTS
Фрейм	FRAME
Фрейм связи	REL_FRAME
Тип связи по указателю	REL_TYPE

В качестве межъязыкового индекса (ILI) используется атрибут «Код синсета» класса «Синсет», а для организации процедур экспорта/импорта с другими проектами на основе WordNet (EuroWordNet, BalkanNet и др.) используется комбинация атрибутов «Офсет синсета» в классе «Синсет» и «Код синтаксической категории» класса «Синтаксическая категория». Таким образом, возможно устанавливать соответствие между синсетами, определенными для английского языка с синсетами, выражающими тот же смысл других языков. Это смысловое соответствие устанавливается экспертами и определяется в толкованиях синсета – для конечных пользователей системы. Сами синсеты и их отношения могут использоваться различными автоматизированными системами и агентами, для которых тезаурус является одной из подсистем.

В случаях, когда в одном из языков нет специальной лексики выражающей значение, представленное синсетом, предлагается для этого языка определять толкование синсета без определения его лексического содержания. Далее, включать такое представление в отношения с другими синсетами, имеющими в своем составе лексические единицы.

Результатом работы с системой является набор тезаурусов WordNet, существующих как самостоятельные базы знаний в рамках отдельно взятого языка и как части тезауруса Princeton WordNet с возможностью организации



функций экспорта/импорта через XML и RDF представления.

Одной из надстроек такой системы является система классификаторов, рис.3. Эта система также является самостоятельной и может быть подсистемой более сложной базы знаний. В частности, расширение и/или подмена второго уровня в классификации WordNet, рис.2 за счет определения соответствия синсетов с разделами классификатора в метасистеме онтологии позволят создать многофункциональную систему определения и классификации знаний.

Рис. 3. UML-диаграмма классов системы классификаторов.

Таблица 3. Наименования классов в модели классификаторов

Наименование	Код
Язык представления	LANGUAGE
Вид описания	DESCTYPE
Вид отношения	LINKTYPE
Категория	CONCEPTTYPE

Наименование	Код
Категория концепта	CONCEPTCLASS
Классификация	CONCEPTCLASSIFICATION
Классификатор	CLASSIFICATION
Концепт	CONCEPT
Представление концепта	CONCEPTLANGUAGE
Описание	CONCEPTDESC
Отношение концепта	CONCEPTRELATION
Форма	CONCEPTFORM

Таблица 4. Наименования атрибутов классов в модели классификаторов

Атрибут	Код
Язык представления	LANG
Язык кратко	LANGSHORT
Дата конца	DATEEND
Дата начала	DATEBEGIN
Вид описания	DESCTYPE
Вид описания кратко	DESCTYPESHORT
Вид отношения	LINKTYPE
Вид отношения кратко	LINKTYPESHORT
Категория	CONCEPTTYPE
Классификатор	CLASSIFICATIONNAME
Классификатор кратко	CLASSIFICATIONSHORTNAME
Код вида описания	DESCTYPEID
Код вида отношения	LINKTYPEID
Код категории	CONCEPTTYPEID
Код классификатора	CLASSIFICATIONID
Код концепта	CONCEPTID
Код отношения	RELATIONID
Код формы	CONCEPTFORMID
Код языка	LANGID
Индекс классификатора	CONCEPTINDEX
Концепт	CONCEPT
Порядок сборки	ITEMORDER
Описание	CONCEPTDESC
Описание в категории	CONCEPTCLASSDESC
Форма	CONCEPTFORM
Форма кратко	CONCEPTFORMSHORT

Краткое описание программной реализации системы

Программа позволяет просматривать, редактировать и добавлять базу данных WordNet. Форма просмотра основной структуры показана на рис. 4 и позволяет работать с любой определенной парой языков, в данном случае – с русским и английским. Интегрированные в нее панели для поиска слов синсета, поиска синсета по фрагменту его толкования и панель рабочей области размещаются и настраиваются по предпочтениям пользователя, что повышает эффективность работы. Они также могут существовать как отдельные окна на рабочем столе и встраиваться в другие формы программы. Для примера, показана форма, в которой осуществлен поиск всех определенных слов, начинающихся с “mag”, из списка слов выбрано “magician”. В области просмотра для этого слова показаны синсеты, они представлены своим толкованием и словарным составом. Для каждого из синсетов также показываются и все синсеты, с которыми у него установлены отношения. Рабочая область содержит ссылки на объекты системы –

синсеты и слова синсетов и используется при редактировании и установлении отношений между синсетами. В рабочей области также присутствуют объекты, не связанные какими-либо отношениями и не до конца определенные, ведется журнал событий, в котором регистрируются все изменения, проводимые пользователем.

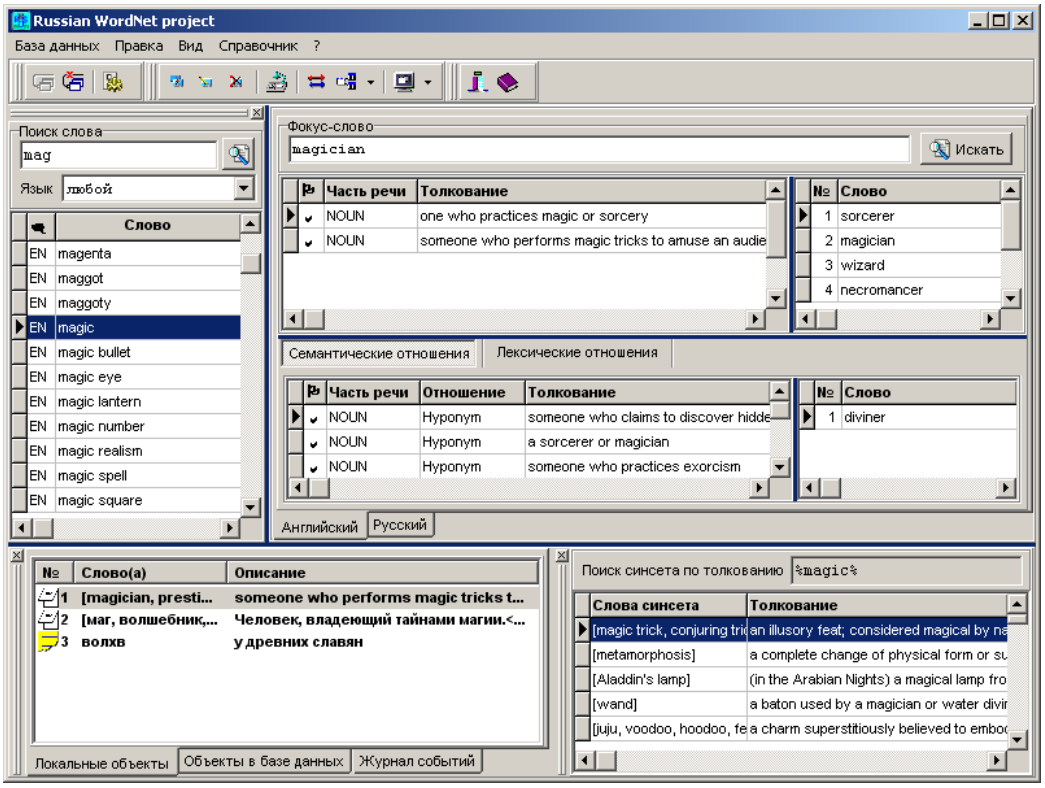


Рис.4. Программа просмотра и редактирования структуры WordNet

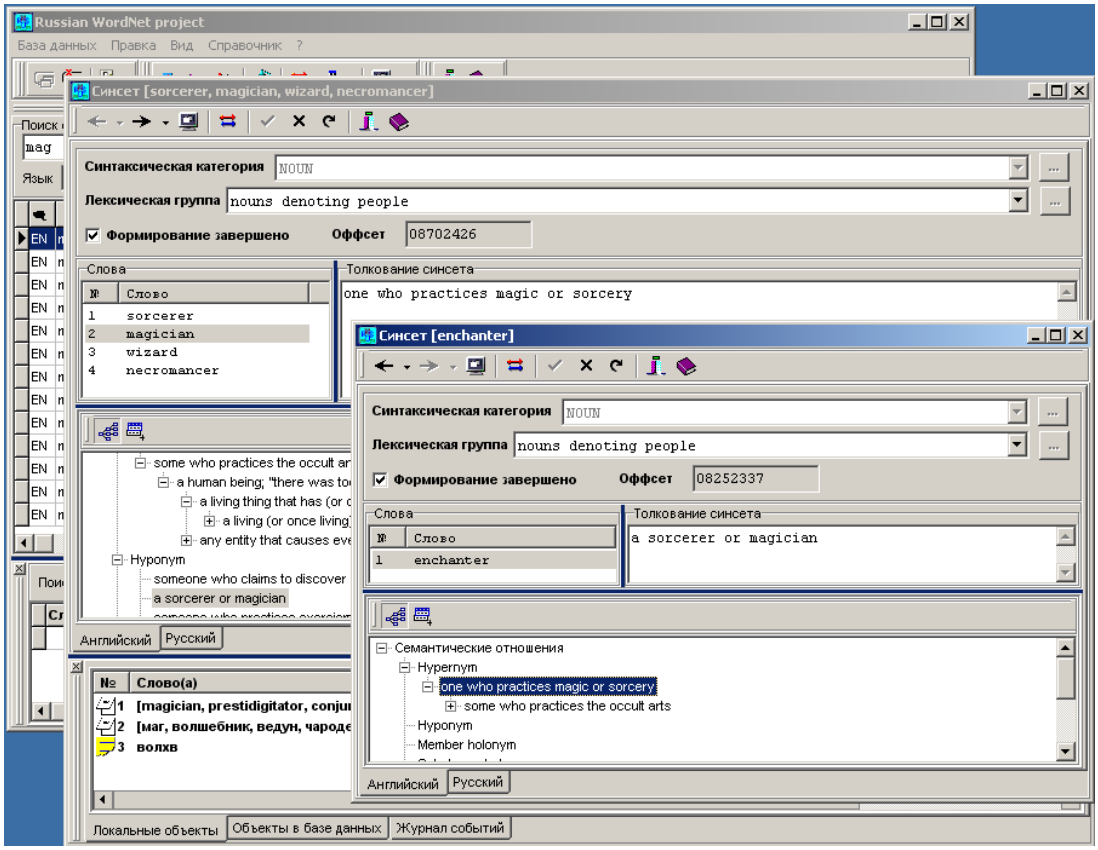


Рис.5. Просмотр и редактирование синсета и его отношений,

Организация формирования и редактирования синсетов и их отношений представлена на рис. 5. Для редактирования каждого синсета организовано отдельное окно. Предоставляется функциональность для редактирования словарного состава синсета, его толкования и отношений. Процесс определения отношений между синсетами облегчается за счет возможностей встраивания панелей (рабочей области, используемой в качестве буфера обмена и других) непосредственно в любую форму редактирования синсета, и за счет организации навигации между всеми формами ввода/редактирования и главной формой.

Для упрощения синхронизации между тезаурусами WordNet, определены функции создания копии синсета и функция объединения двух произвольных синсетов, рис. 6.

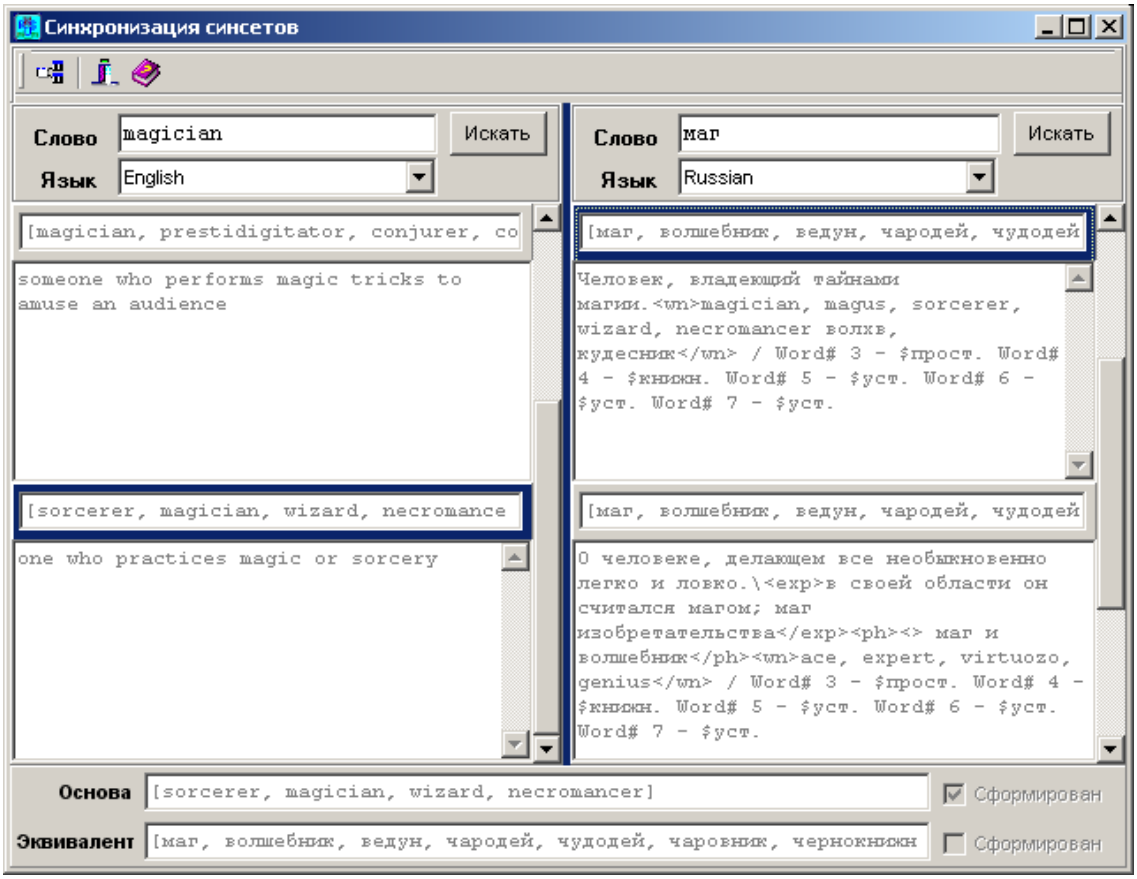


Рис.6. Синхронизация синсетов

Дополнительно к системе подключается морфологический анализатор, корпус текстов и множество электронных словарей Руссикон [6]:

- грамматические (общей и предметной лексики);
- орфографический;
- толковый;
- англо-русский и русско-английский.

Заключение

Рассмотренные системы предназначаются для создания и редактирования широкого класса тезаурусов и близких к ним структур. Реализация набора интерфейсов к этим системам позволяет использовать их как самостоятельные приложения – лексикографическая система WordNet и система классификаторов, так и включать их в состав более сложных систем. Планируется расширение таким образом уровня лексических категорий WordNet классификациями, определенными отечественными стандартами и проверенными на практике в библиотечно-издательском деле.

Список литературы

1. WordNet. An Electronic Lexical Database. Christiane Fellbaum (ed.). Bradford Books.
2. Bernardo Magnini and Gabriela Cavaglia'. Integrating Subject Field Codes into WordNet. In: Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and

Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 MAY- 2 JUNE 2000, pp. 1413-1418.
3. A Discussion of the Relationship Between RDF-Schema and UML. W3C Note 04-Aug-1998, http://www.w3.org/TR/NOTE-rdf-uml/.
4. Booch, G., Rumbaugh, J., and Jacobson, I., 1998. The Unified Modeling Language user guide, Addison-Wesley.Didier M., et al., 2000. Professional XML. Wrox Press Ltd.
5. Prószéky, Gábor & Márton Miháltz, 2002. Semiautomatic Development of the Hungarian Word-Net.LREC-2002, Las Palmas, Spain.
6. Yablonsky S.A., 1998. Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources & Evaluation, Granada, Spain.
7. Yablonsky S. A., 2002. Corpora as Object-Oriented System. From UML-notation to Implementation. In: Proceedings Third International Conference on Language Resources & Evaluation LREC-2002, Las Palmas, Spain.
8. ГОСТ 7.49-84 Система стандартов по библиотечному и издательскому делу. Рубрикатор ГАСНТИ. Структура, правила пользования и ведение.
9. ГОСТ 7.25-2001 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.